Calculus for Mathematicians, Computer Scientists, and Physicists

An Introduction to Abstract Mathematics

Andrew D. Hwang

Contents

List of Figures

Preface

Calculus is an important part of the intellectual tradition handed down to us by the Ancient Greeks. Refined in the intervening centuries, calculus has been used to resolve a truly incredible array of problems in physics, and is increasingly applied in areas as diverse as computer science, population biology, and economics. As taught in the schools, "calculus" refers almost exclusively to the calculational procedures—differentiation and integration in one real variable—used in the mathematical study of rates of change. However, calculus has a beautiful but lesser-known theoretical foundation that allows us to speak consistently and meaningfully of the infinitely large and infinitesimally small. This foundation, which is required knowledge for all serious students of mathematics, exemplifies the dual aspects of theory and application in mathematics.

The aims of this book are various, but all stem from the author's wish to present beautiful, interesting, living mathematics, as intuitively and informally as possible, without compromising logical rigor. Naturally, you will solidify your calculational knowledge, for this is in most applications the skill of primary importance. Second, you will acquire understanding of the theoretical underpinnings of the calculus, essentially from first principles. If you enjoy pondering the concept of infinity, this aspect of the book should appeal to you. The third major goal is to teach you something of the nature and philosophy of *mathematics itself*. This aspect of the presentation is intended to have general appeal, not just to students who intend to major in mathematics at the university level, but to the mathematically curious public. Calculus is extremely well-suited to this meta-lesson, because its theoretical foundations rest firmly upon notions of infinity, which can lead to apparent logical paradoxes if not developed carefully. To make an analogy, trying to apprehend the logical nature of calculus by intuition alone is akin to landing an airplane in cloudy weather; you may succeed if someone

trustworthy tells you what to do (or if an author states theorems without proof, or even without giving careful definitions), but you acquire few flying skills that can be applied at other airports (or in other mathematical situations). Careful definitions, logic, and proof are the radar that allow you to see through the intuitive fog, resolve (or avoid) apparent contradictions, and understand what is true and *why* it is true. When you have mastered the organization of ideas required to prove a theorem, the theorem becomes *a part of you* that cannot be taken away or denied by anyone. By contrast, factual knowledge acquired by memorization is shallow and flimsy: If you are told that something you have learned is wrong, you have no way to get at the truth. To continue the piloting analogy, if you only know how to fly when an instructor is telling you what to do, your skills are not generally applicable. The information the radar gives you (statements of theorems) is important, but even more important is your ability to use radar for yourself, especially in new situations.

The fourth and final large goal of the book is to present substantial mathematical results. Despite common perception, mathematics is a creative discipline, often likened to classical music by its practitioners. Just as no one would mistake musical scales for actual music, nor would anyone confuse spelling and grammar for literature, no mathematician would equate routine book problems with real mathematics. Of course, routine problems are an important pedagogical tool: They allow you to practice computational techniques until you are fluent. But while you are not likely to succeed mathematically unless you master calculation, you are certainly not guaranteed success merely by working lots of routine problems. If your only experience with mathematics is school courses, you may have much to unlearn about the nature of mathematics. The material you encounter may seem qualitatively unfamiliar at first, but gradually your viewpoint will shift away from techniques towards the concepts and logical relationships that are fundamental to the nature of mathematics. Along the way, you will also meet colorful identities such as

$$e^{i\pi} + 1 = 0 \qquad \text{and} \qquad \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6},$$

and precisely stated assertions that encapsulate facts you learned in school, but with terms carefully defined and logical structure laid out plainly.

To make one last analogy, think of mathematics as terrain. Some areas are flat and grassy, good for a pleasant stroll; others are hilly, but with well-worn trails. There are cliffs scaled by narrow, difficult paths that require ropes, and ravines where it is easy to see how to cross, but difficult to scramble through the undergrowth. Here and there, mountain peaks climb far above the plains. The views from the top are stunning, allowing you to see wide vistas and their connections, but are achieved only by a long, difficult climb. Still, the trails have been blazed, and steady work will bring you to the summit. The book is your tour guide, showing you interesting trails and geographical features in the part of the map called "real analysis". The skills you learn will accumulate, allowing you to tackle more difficult trails. Along the way, there are springs, waterfalls, and wildflowers as rewards for steady progress. All the places on the itinerary are well-mapped out, but occasionally we will pass a cliff that no one has yet scaled, and a few times you will glimpse the vast, uncharted hinterlands where current research occurs.

Chapter 1

The Language of Mathematics

> Die Mathematiker sind eine Art Franzosen; redet man zu ihnen, so übersetzen sie es in ihre Sprache, und alsbald ist es etwas ganz anderes.
>
> *Mathematicians are like Frenchmen. Whatever you say to them, they will translate into their own language, whereupon it becomes something completely different.* —Goethe

1.1 The Nature of Mathematics

Mathematics is unique among human intellectual endeavors; it is not art, philosophy, or science, but it shares certain features with each. As a language, mathematics is unparalleled in its ability to express features of the physical world with astounding accuracy. At the same time, mathematics has no intrinsic connection to the real world; the objects of mathematics are concepts, and do not have physical existence in the same way that stars, molecules, or people do. Conversely, stars, molecules, and people are not mathematical objects, though they do possess attributes (height, mass, or temperature, for example) that can be modeled by mathematical concepts.

It is remarkable that the language of mathematics can be used to describe and predict natural phenomena, but this fact is not itself part of mathematics. Mathematical concepts seem to exist independently, in a (metaphorical) Platonic universe apart from the physical world, waiting to be *discovered* rather than *created*. Mathematics is guided to an extent by aesthetics, and though there are no graduate-level courses on what constitutes beautiful mathematics, mathematicians agree to a

remarkable extent in judging parts of mathematics to be beautiful, or deep (meaning "connected to widely disparate parts of the subject"), or correct (meaning "logically consistent" rather than "true").

Electronic computers provide a good physical analogy for *abstraction* and *recursion*, two basic themes of this book. Briefly, abstraction is the process of describing the properties an object or structure possesses. Recursion is the act of defining a structure in terms of simpler objects. Abstraction is in contrast to *implementation*, which is a specific way of building or realizing something.

For example, a computer stores and manipulates data as a pattern of *bits* or *binary digits*, conventionally called 0 and 1. Bits are an abstraction of data, and abstract properties of data are those that are independent of the physical representation of bits. In many theoretical considerations, only the pattern of bits, the abstract structure of the data, is relevant.

The recursive nature of computer science is exemplified by the fact that a computer program is not merely a long string of bits, but that these bits are organized, first into groups of "bytes" (almost always 8 bits), then into "words" (of 4 or 8 bytes), then assembly language instructions (collections of words that are meaningful to the processor), functions (groups of assembly instructions that perform some action specified by the programmer), and libraries (collections of related functions). A computer program—your web browser, text editor, or media player—is built by "linking" functions from libraries.

Data storage can be implemented physically in several ways:

- Magnetic domains—floppy and ZIP disks, PC hard drives.
- Light and dark spots or bands—compact disks and UPC symbols.
- Holes and "no-holes" in a paper strip—punch cards and ticker tape.
- Charged and uncharged capacitors—RAM.

The common abstract feature is a *pair of contrasting states*. A mathematician or theoretical computer scientist sees no essential difference between these storage schemes. Depending on context, we might call the contrasting states "black and white", "zero and one", "true and false", or "on and off", but regardless of name there is a single underlying structure. In mathematics, you should strive to understand the structure rather than memorize the name.

Next, consider the following arithmetic/logical operations that are built from representations of bits:

- (Binary arithmetic) Think of 0 as representing an arbitrary even integer, and 1 as representing an arbitrary odd integer. That is, an integer is identified with its remainder on division by 2. The sum of two odd integers is always even ("$1+1=0$"), the product of an even and an odd integer is always even ("$0 \cdot 1 = 0$"), and so forth. If we tabulate the results of addition and multiplication, we get

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

- (Boolean logic) Think of F as representing an arbitrary "false" assertion (such as "$2+2=5$") and T as representing an arbitrary "true" sentence (such as "$1+1=2$"). Since "2+2 = 5 or 1+1 = 2, but not both" is true, we write "F xor T=T". ("xor" stands for "exclusive or": one statement is true, but not both.) Since "2+2 = 5 and 1+1 = 2" is false, we write "F and T=F". The tables below give the "truth value" of a statement made by conjoining two statements, according to whether or not each statement is true or false.

xor	F	T
F	F	T
T	T	F

and	F	T
F	F	F
T	F	T

Each pair of tables encapsulates some structure about bits of data. The truly mathematical observation is that *the entries of the tables correspond*: under the correspondence even-False and odd-True, "addition (mod 2)" corresponds to "xor", and "multiplication (mod 2)" corresponds to "and". The two pairs of tables above are different implementations of the same abstract structure, which might even be denoted

$\vee$	$\bullet$	$\circ$
$\bullet$	$\bullet$	$\circ$
$\circ$	$\circ$	$\bullet$

$\wedge$	$\bullet$	$\circ$
$\bullet$	$\bullet$	$\bullet$
$\circ$	$\bullet$	$\circ$

We have just described an abstract relationship among abstract structures. Eventually, we might find such relationships arising in various contexts and give the concept a name, such as "isomorphism". Making such a definition is an act of recursion: It groups together a class of abstract relationships among abstract structures. In time, we might find it profitable to study isomorphisms themselves, moving to yet a higher level of abstraction. The branch of mathematics called *category theory* deals with this sort of issue.

The example above is intended not to discourage you with complexity, but to illustrate that mathematics is concerned with abstract properties of conceptual systems, and that structures can be organized and understood with their extraneous details stripped away.

The Objects of Mathematics

The fundamental objects in mathematics are *sets* and their constituent *elements*. A set is an abstraction of the informal concept of a collection of objects. The set of names of the U.S. states in the year 1976 has 50 elements; "Massachusetts" is an element of this set, while "Puerto Rico" and "Uranium" are not. As a more mathematical example, the set of *prime numbers*, integers p greater than 1 that have no divisors other than 1 and p, is a set. The numbers 2, 5, and $2^{13466917} - 1$ are elements of the set of primes, while 4 and $2^{13466917}$ are not.

In mathematics, we do not discuss "what sets really are"; this issue is philosophical, not mathematical. Mathematics is built on set theory in much the same way computer science is built on strings of bits. In computer science, the objects of study are built up recursively; ultimately, everything is defined in terms of Boolean operations and words of bits. Further, a computer program has an abstract existence aside from the way the bits are stored and accessed physically. Similarly, the objects and structures of calculus—integers, real numbers, functions, and so forth—are defined recursively in terms of simpler objects, and are ultimately built from sets. Every mathematical assertion may be interpreted as an assertion about sets, though even "$2 + 2 = 4$" is surprisingly difficult to write purely in terms of sets. The nature of sets is an irrelevant "implementation detail"; instead, the *properties* of sets are paramount.

Mathematics and Science

Mathematics was once called "the most exact science". In the last two centuries, it has become fairly clear that mathematics is fundamentally not a science at all. In mathematics, the standard of acceptance of an idea is logical, deductive proof, about which we say more below. While mathematicians sometimes perform "experiments" , either with pencil and paper, or with an electronic computer, the results of a mathematical experiment are never regarded as definitive. In physics or chemistry, by contrast, experiment is the sole criterion for validity of an idea.[1]

A few minutes' reflection should reveal the reasons for these radically differing criteria. Mathematical concepts have no relevant attributes other than those we ascribe to them, so in principle a mathematician has complete access to all properties possessed by an object. In the physical sciences, however, the objects of study are phenomena, about which information can only be obtained by experiment. No matter how many experiments are performed, scientists can never be certain that their knowledge is complete; a more refined experiment may conflict with existing results, indicating (at best) that an accepted Law of Nature needs to be modified, or (at worst) that someone has collected data carelessly. Experimental results are never mathematically exact, but are subject to "uncertainty" or "experimental error". Thus, in the sciences we do not have the same access to our object of study that we do in mathematics. Laws of Nature—mathematical models of some aspect of reality—are virtually assured of being approximate.

Despite the differing aims and standards of acceptance, mathematics and the physical sciences enrich each other considerably. The most obvious direction of influence is from mathematics to the sciences: The best available descriptions of natural phenomena are mathematical, and are astoundingly accurate. For example, total eclipses of the sun can be predicted hundreds of years in advance, down to the time and locations at which totality will occur. Less apparent but no less important is the beneficial influence that physics, biology, and economics have had on mathematics, particularly in the 20th Century. For whatever reason, mathematics that describes natural phenomena is deeply interconnected and full of beautiful, unexpected results. Without the guiding influence of science, mathematics tends to become ingrown, specialized, and merely technical.

[1] This characterization of science is due to the physicist, R. P. Feynmann.

Mathematical Certainty

This chapter is informal, and its intent is to make contact with material you know rather than lay down formal foundations of mathematics. Nonetheless, formal foundations do exist, in the form of sets of axioms of set theory. The "usual" axioms are called *ZFC*, for Zermelo-Frankel and the axiom of Choice.

In mathematics, there is no concept of absolute truth, only *logical consistency* with respect to an axiom system, about which we say more below. It is an article of faith among mathematicians that ZFC is consistent. Most mathematicians work within ZFC, and results that are proved in ZFC are said (colloquially) to be "true". However, it is important to remember that mathematics does not really produce "objective truth" but rather establishes that statements are consistent with ZFC. The distinction is fundamental, and often misunderstood. To say that "$2 + 2 = 4$ is a universal, mathematical truth" is misguided; more accurate would be the legalistic (yet non-trivial) claim, "If the concepts 2, 4, $+$, and $=$ are suitably defined, in a way that conforms to our intuition about counting, then $2 + 2 = 4$." Mathematical truth, arguably the most certain kind of truth, is always relative to an axiom system. Provability, not truth itself, is the central concern of mathematics.

1.2 Sets and Operations

Sets are often denoted by capital letters, and elements by small letters. We write $x \in X$ to indicate that X is a set and x is an element of X. Similarly, we write $y \notin X$ to denote the fact that y is not an element of X.

If Y is a set with the property that every element of Y is an element of X, then we say Y is a *subset* of X and write $Y \subset X$. The names of the original thirteen colonies are a subset of the set of state names. The set of prime numbers is a subset of the set of positive integers. The set of even numbers is *not* a subset of the set of prime numbers.

In order to avoid logical contradictions (such as *Russell's paradox*, see Exercise 1.5), it is necessary to fix a *universe*, a set $\mathcal{U}$ with the property that $X \subset \mathcal{U}$ for every set X under consideration. In this book the universe is usually taken to be $\mathbf{R}$, the set of real numbers, or the set of "functions" whose domain and range are $\mathbf{R}$. (Functions are defined in Chapter 3.)

Sets can be visualized with *Venn diagrams*. The universe is depicted as a rectangle, and sets under consideration are interiors of curves, Figure 1.1.

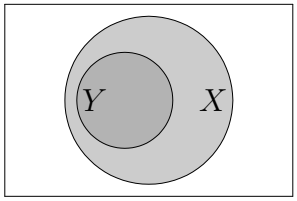


Figure 1.1: A Venn diagram for the relation $Y \subset X$.

Two sets X and Y are *equal* exactly when $X \subset Y$ and $Y \subset X$, namely, when each set has the same elements. A set that has finitely many elements may be presented as a list, as in $\{0,\ 1\}$. It does not matter if an element is listed more than once; the sets $\{0,\ 1\}$ and $\{0,\ 1,\ 1\}$ are equal. "Set builder notation" describes a set by specifying the universe together with properties that characterize the elements. For example, $\mathbf{R}^+ := \{x \in \mathbf{R} \mid x > 0\}$ (read "the set of x in $\mathbf{R}$ such that x is greater than 0") denotes the set of positive real numbers. The colon next to the equality indicates that the corresponding side of the equation is being defined.

There is a unique *empty set* $\varnothing$ which contains *no* elements. If you have trouble believing this, you will sympathize with medieval philosophers who balked at the concept of "zero." However, the deciding factors are that the empty set is useful, and its existence is logically consistent with the axioms of set theory.

An element x of a set X should be carefully distinguished from the *singleton* $\{x\}$, the set whose single element is x. For example, the relation $x \in \{x\}$ is always true, while $x \in x$ is rarely true. Similarly, elements and subsets must not be confused, though on the surface they are related concepts: For every set X, we have $\varnothing \subset X$ but usually $\varnothing \notin X$.

A set can be specified in many ways; for instance, $\mathbf{R}^+$ is also expressed as $\{y \in \mathbf{R} \mid y \text{ has a real logarithm}\}$, or as $\{x \in \mathbf{R} \mid x \neq 0 \text{ and } x = \sqrt{x^2}\}$, while the two-element set $\{0,\ 1\}$ can be written as $\{x \in \mathbf{R} \mid x^2 = x\}$, for example.

Specifications of the empty set are often amusing:

$$\varnothing = \{x \in \mathbf{R} \mid x \neq x\} = \{x \in \mathbf{R} \mid x^2 = -1\} = \{x \in \mathbf{Q} \mid x^2 = 2\}.$$

The last characterization is discussed below; "$\mathbf{Q}$" is the set of rational numbers. The empty set must be distinguished from the (non-empty) set $\{\varnothing\}$, whose single element is the empty set. This point is not at all silly, despite appearances, see Chapter 2.

Denote the universe by $\mathcal{U}$. If $X \subset \mathcal{U}$ is a set, then the *complement* of X, denoted $\sim X$ or X^c, is the set defined by $X^c = \{y \in \mathcal{U} \mid y \notin X\}$. Informally, the complement of X is "the set of objects not in X".

There are operations from which new sets are formed from existing sets. Four of the most important are informally described here. If X and Y are sets, then we may form their

- *Union* $X \cup Y$, consisting of elements that are in X or Y or both. (Mathematicians use the word "or" in the non-exclusive sense unless specifically stated otherwise.)

- *Intersection* $X \cap Y$, consisting of elements that are in both X and Y. The sets X and Y are *disjoint* if their intersection is the empty set, that is, they have no elements in common.

- *Cartesian product* $X \times Y$, consisting of all *ordered pairs* (x, y) with $x \in X$ and $y \in Y$. If $X = \{A,\ B,\ \ldots, H\}$ and $Y = \{1,\ 2,\ \ldots,\ 8\}$, then $X \times Y$ is the 64-element set

  $$\{(A,1),\ (A,2),\ \ldots,\ (A,8),\ (B,1),\ \ldots,\ (H,8)\},$$

  while if X and Y are intervals of real numbers, then their Cartesian product is a rectangle in the plane. In particular, a plane may be viewed as the Cartesian product of two lines.

- *Difference* $X \setminus Y$, consisting of all elements of X that are *not* elements of Y. If the universe is fixed, then $X \setminus Y = X \cap (\sim Y)$.

Union and intersection are Boolean operations ("or" and "and" respectively), while the Cartesian product creates tables from lists. Venn diagrams for union, intersection, and difference are as follows:

Unions and intersections of infinitely many sets are defined as expected: If $\{X_\alpha\}_{\alpha \in I}$ is a family of sets in some universe $\mathcal{U}$ (with I an "index set"), then

$$\bigcup_{\alpha \in I} X_\alpha = \{x \in \mathcal{U} \mid x \in X_\alpha \text{ for } \textit{some } \alpha \in I\},$$
$$\bigcap_{\alpha \in I} X_\alpha = \{x \in \mathcal{U} \mid x \in X_\alpha \text{ for } \textit{all } \alpha \in I\}.$$

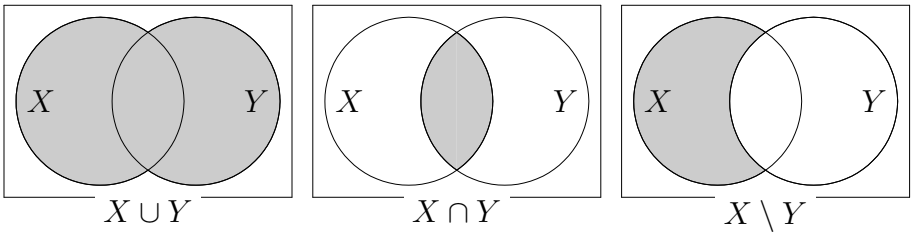


Figure 1.2: Venn diagrams for simple set operations.

In this book, an infinite index set is usually the set of positive integers, as in

$$X = \bigcap_{n=1}^{\infty} (-\tfrac{1}{n}, \tfrac{1}{n}).$$

A real number x is an element of X exactly when $-1/n < x < 1/n$ for every positive integer n. We will see that $x = 0$ is the only real number with this property, so $X = \{0\}$. Additional examples of infinite set operations are given in Exercise 2.10.

1.3 Logic

As mentioned, mathematicians do not really find "objective truths", but instead derive *logical conclusions* by starting with assumptions called *hypotheses*. In this section we begin the study of logical deduction, emphasizing the linguistic differences with ordinary English.

Mathematicians use generally agreed-upon *axioms* (foundational assumptions) and rules of *logical deduction*. In this book, the axioms are the *Zermelo-Frankel-Choice* axioms of set theory, and the rules of deduction are those of *Aristotelean logic*.

Throughout this section, a running sequence of examples is presented to illustrate the concepts being introduced. Mathematics aims for generality, but the human mind often favors particulars, and it is these that make mathematics directly useful. The aim of mathematics is precise thinking, not generality for its own sake. That said, abstraction (which fosters generality) has a definite purpose: To extract the essential features of a problem and ignore extraneous details. Precision is important because intuition (especially regarding the infinite) is often misleading, sometimes blatantly wrong. Logical deduction is

the "hygiene of mathematics," (after H. Weyl) the principle tool by which intuition is checked for logical consistency, and erroneous thinking avoided.

Statements and Implications

A *statement* is a sentence that has a *truth value*, that is, which is unambiguously either true or false with respect to some axiom system. A mathematical sentence may depend on *variables*, and may thereby summarize a family of statements, one for each possible assignment of variables. The truth value of the resulting statement may depend on the values of the variables. The important thing is that, for each specific choice of variables, the sentence should be either true or false. By use of variables, theorems often encapsulate infinitely many statements.

Here are some examples, in which the variable n is an integer. Observe that the statements that depend on n encapsulate infinitely many statements.

- "-4 is an even integer." "The decimal expansion of π is non-repeating and contains the string '999999'." (True)
- "For every integer n, $n^2 - n$ is an even integer." (True)
- "$2 + 2 = 5$." (False)
- "For some integer n, both n and $n + 1$ are even integers. (False)

Sentences that are *not* statements include "n is an even integer" (whose truth value depends on n) and the self-referential examples, "This sentence is true" (whose truth value must be specified as an axiom) and "This sentence is false" (which cannot be consistently assigned a truth value).

Statements are linked by *logical implications* or *if-then statements*, sentences of the type, "If H, then C." The variable H is a statement, called the *hypothesis* of the implication, and the variable C is a statement called the *conclusion* of the implication. We think of C as being *deduced* or *derived* from H.

The fundamental idea of Aristotelean logic is that an implication is *valid* unless it derives a false statement from a true statement:

- If $1 \neq 0$, then $1^2 \neq 0$. (Valid)

- If $1 \neq 0$, then $1^2 = 0$. (Invalid)
- If $1 = 0$, then $1^2 \neq 0$. (Valid)
- If $1 = 0$, then $1^2 = 0$. (Valid)

If a hypothesis and conclusion are related by valid implication, then the hypothesis is said to *imply* the conclusion. In this view, it is valid (not logically erroneous) to deduce a conclusion from a false hypothesis: If we start with truths and make valid deductions, we obtain only truths, not falsehoods. An implication with false hypothesis is said to be *vacuous*. To emphasize, validity has the possibly counterintuitive property that if the hypothesis is false, then *every* conclusion follows by valid implication. As strange as this convention seems, it does not allow us to deduce falsehoods from truths. Logical validity is central to the concept of "proof," and is therefore crucial to the rest of the book (and to mathematics in general).

The term "imply" has a very different meaning in logic than in ordinary English. In English, to "imply" is to "hint" or "suggest" or "insinuate." In mathematics, if a hypothesis implies a conclusion, then the truth of the conclusion is an ironclad certainty provided the hypothesis is true. The term "valid" also has a precise meaning which is not exactly the same as in ordinary English. Finally, note that every statement has a truth value, but only an if-then statement can be valid or invalid.

Interesting logical implications usually depend on variables, and the truth value of the implication therefore depends upon the truth values of the hypothesis and conclusion. The concept of logical validity comes into its own when an implication depends on variables. The following examples illustrate various combinations of truth and falsehood in hypothesis and conclusion:

- If n is an integer and if n is even, then $n/2$ is an integer. (Valid)
- If n is an integer, then $n/2$ is an integer. (Invalid)
- If n is an integer and $n = n + 1$, then $2 + 2 = 5$. (Valid)
- If n is an integer and $n = n + 1$, then $2 + 2 = 4$. (Valid)

The distinction between "truth" (which applies to statements) and "validity" (which applies to logical implications) may at first seem a bit fussy. However, it is important to be aware that the concepts are different, though they are not wholly unrelated, either; when a logical

implication has a definite hypothesis and conclusion, the *entire sentence* becomes a statement, which may be either true or false. Validity of the implication expresses the truth of the resulting statement in terms of the truth values of the hypothesis and conclusion. In this sense, logical validity is a sort of "meta-truth." A single valid implication usually yields infinitely many true statements.

Logical Consistency

If, in some axiomatic system, it is possible to prove some proposition P and also prove the negation $\neg P$, then *every* statement Q is provable, since either $P \implies Q$ or $\neg P \implies Q$ is vacuously true. The pair $\{P, \neg P\}$ is called a *logical contradiction*, and an axiom system is said to be *inconsistent* if a contradiction can be derived in it. While it may be very interesting to discover that an axiom system is inconsistent, an inconsistent system is not in itself mathematically interesting.

Work of K. Gödel in the 1930s showed that it is impossible to prove that ZFC is consistent, except by using some other ("more powerful") axiom system, whose consistency is unknown. By way of reassurance, it is also known that if there is a contradiction in ZFC, then there is a contradiction in ordinary arithmetic.

Definitions and Theorems

Mathematical definitions establish terminology, the common ground from which to work. The primary difficulty in making "good" definitions is isolating, or *abstracting*, exactly the desired conceptual properties.

Mathematical definitions are interpreted literally. For a beginner, it can be a serious conceptual obstacle not to read in more than is stated when interpreting definitions.

> A physicist, a statistician, and a mathematician were motoring in the Scottish countryside when they came upon a flock of one hundred sheep, one of which was black. The physicist said, "From this, we deduce that one percent of sheep are black." The statistician said, "No, we only know that of *these* 100 sheep, one is black." The mathematician corrected, "I'm afraid you're both wrong. We only know that of these 100 sheep, one of them is black *on one side*."

When you are asked to prove something (in this book or elsewhere), the first thing to do is make sure you know and understand the definitions of all concepts involved. Eventually this will become second nature to you (or you will quit mathematics in frustration), but it doesn't hurt to be reminded frequently at this stage.

Definition 1.1 An integer n is *even* if there exists an integer m such that $n = 2m$.

For each integer n, the definition provides a criterion to determine whether or not n is even. The criterion is a pass/fail test, nothing more. A definition also provides a *condition* that every even integer satisfies. If an integer k is even, you immediately know something about k; for instance, the last digit cannot be "3."

We immediately see that $n = 6$ is even, since $m = 3$ satisfies the condition of the definition. By contrast, we cannot determine so easily whether 5 is even or not; using the definition, we would have to show somehow that $5 \neq 2m$ for every integer m, an infinite task. Instead we must find a criterion for non-evenness. For example, we might prove that an integer that leaves a remainder of 1 on division by 2 is not even. Since 5 satisfies this criterion, we would deduce that 5 is not even.

Generally, in problem solving the definition plays the role of a test, e.g. "Determine which of the following integers is even..." while in proving theorems, the definition plays the role of a condition, e.g. "If n is an even integer, then n^2 is an even integer."

On Doing Mathematics

The dilemma above ("Is 5 even?") is the norm for mathematical study and research, and can be extremely frustrating. A mathematical criterion is usually considerably more subtle than "evenness", and it may be difficult to see immediately whether or not a specific object satisfies a criterion. Unfortunately, if you do not know in advance what is true, you do not how to proceed when trying to prove something! All too quickly you will encounter this survival lesson: In mathematics, you must not be afraid to work provisionally, follow blind alleys, examine the consequences of hypotheses not known to be true, and search for examples that may not exist. The process of discovery is never straightforward, and mathematics is no exception. In time you will develop intuition regarding approaches to a problem that are likely to be fruitful, and ideas that are probably dead ends. You will learn how to "play" with hypotheses, how to look at special cases and formulate

general guesses, how to distinguish real patterns from illusory ones, and finally how to prove that the patterns you have found *are* real.

Theorems

A *theorem* is a valid implication of sufficient interest to warrant special attention. A *lemma* is a valid implication that is mostly of technical interest in establishing a theorem. If you program, it may help to think of a lemma as a "logical subroutine", a short piece of logical argument that is used repeatedly and should be separated out for clarity and brevity. A *proposition* is "a small theorem of independent interest". To an extent, the choice of term in a given situation is a matter of style.

Most mathematical assertions are stated in one of three forms (in roughly decreasing order of formality):

- (Valid implication) "If n is an even integer, then n^2 is a multiple of 4."
- (Quantified sentence) "For every even integer n, n^2 is a multiple of 4."
- (Direct statement) "The square of an even integer is a multiple of 4."

Each expresses the fact that an object that has one property (an even integer) also has another property (its square is a multiple of 4). An even more formal wording, that combines implication and quantification, is "If n is an even integer, then there exists an integer k such that $n^2 = 4k$."

When *every* hypothesis of a valid implication is true, then the conclusion is also true. This is the only information implicitly or explicitly conveyed by a logical implication. In particular, if some hypothesis is false, then no information whatsoever is asserted. To emphasize:

> *A theorem conveys absolutely no information unless* ***every*** *hypothesis is satisfied.*

A common confusion is to remember the conclusion of a theorem and to pay no attention to the hypotheses, thereby leading (at best) to a statement out of context or (at worst) to a bad interpretation. In the mid-1990s, a popular newspaper columnist fell into this pitfall

over A. Wiles' proof of "Fermat's Last Theorem". Wiles used techniques from "hyperbolic geometry", which the columnist thought was self-contradictory because "It is possible to square the circle in hyperbolic geometry", while every school student of the columnist's generation learned that "It is impossible to square the circle". The columnist was, presumably, remembering the conclusion of a celebrated theorem of 19th Century mathematics:

Theorem 1.2. *Let the axioms of Euclidean geometry be assumed. If a line segment of unit length is given, then it is impossible to construct a line segment of length π in finitely many steps using only a straightedge and compass. Consequently, it is impossible to construct a segment of length $\sqrt{\pi}$, that is, to "square the circle".*

As a general lesson, Theorem 1.2 says nothing about the possibility of constructing such a line segment with tools other than a straightedge and compass, nor about the possibility of obtaining better and better approximations with a straightedge and compass, thereby (in a sense) achieving the construction in infinitely many steps. The relevant shortcoming in this story is that the theorem says nothing unless the axioms of Euclidean geometry are assumed.[2]

If there is a linguistic lesson to be gleaned from mathematics, it is that words themselves are merely labels for concepts. While our minds react strongly to words,[3] it is the underlying *concepts* that are central to logic, mathematics, and reality. Good terminology is chosen to reflect meaning, but it is a common, *human*, mistake to assume an implication is obvious on the basis of terminology. Mathematicians remind themselves of this with the *red herring principle*:

> In mathematics, a 'red herring' may be neither red nor a herring.

Theorem 1.2 is remarkable for another reason: It asserts the *impossibility* of a procedure that is *a priori* conceivable (namely, that is not

[2]The columnist's error was not this glaring; they argued that because theorems of hyperbolic geometry can be interpreted as statements in Euclidean geometry, a "hyperbolic" proof is self-contradictory. The resolution to this objection is that while "squaring the circle in hyperbolic geometry" can be interpreted as a statement about Euclidean geometry, the interpretation is markedly different from "squaring the circle in Euclidean geometry", and does not contradict Theorem 1.2.

[3]To the extent that nonsensical rhetoric can be persuasive, or that it is illegal in the U.S. to broadcast certain words by radio or television.

obviously contradictory).[4] This is a completely different matter from saying "Human knowledge does not currently have the means to 'square the circle'." It means, rather, that the axioms of Euclidean geometry *are not logically compatible* with the construction of a certain line segment, with certain tools, in finitely many steps. The logic of the proof is briefly sketched later on, after methods of proof have been discussed.

Proof

To illustrate the ideas of proof in more detail, and with an example of some historical and mathematical importance, consider the familiar fact that "$\sqrt{2}$ is irrational." There is a theorem behind this assertion, but the present phrasing leaves much to be desired: It ignores, for example, questions such as "What is a real number?" and "What is the relationship between rational and irrational (real) numbers?" A fastidious mathematician might prefer the assertion "There is no rational square root of 2." Even this is not a logical implication, however. One way to express precisely (and in a manner amenable to proof!) the irrationality of $\sqrt{2}$ is as follows.

Theorem 1.3. *If m and n are positive integers, then $(m/n)^2 \neq 2$.*

Theorem 1.3 exemplifies the quote of Goethe at the opening of this chapter, though with a bit of practice you will be able to translate mentally from informal assertions to precisely stated logical implications. You should convince yourself this theorem really does say "There is no rational square root of 2." An alternative wording is the quantified sentence, "For every rational number x, $x^2 \neq 2$."

The proof of Theorem 1.3 is expedited by the following observation:

Lemma 1.4. *If k is an even integer, and if there is an integer m with $m^2 = k$, then k is a multiple of 4.*

In Lemma 1.4, the hypothesis consists of two statements, "k is an even integer" and "there is an integer m with $m^2 = k$" (sometimes phrased as "k is a square"). The conclusion is the statement "there is an integer n such that $k = 4n$." As stated, Lemma 1.4 gives no information whatsoever in the event that k is not even, nor does it give information if k is not a perfect square.

[4] A popular cartoonist claimed that "It is impossible to prove the impossibility of something." While this is arguably true of science, it is certainly not true of mathematics.

Proof. Lemma 1.4 is established by checking a couple of cases. Assume that k is a square, so there is an integer m with $m^2 = k$. If m is odd, then $k = m^2$ is odd (why?), while if m is even, then $k = m^2$ is a multiple of 4 (why?). So, the only way a square can be even is if it is already a multiple of 4, as was to be shown. □

This proof is admittedly a bit informal, if conceptually correct; the details would require definition of an "odd" integer, together with steps answering the two questions in parentheses. A direct proof of Theorem 1.3 can be based on Lemma 1.4. A more standard proof by contradiction is given later.

Proof. Observe that $(m/n)^2 = 2$ means the same thing as $m^2 = 2n^2$. By writing m/n in lowest terms, m and n may be assumed to have no common factor; in particular, they are not both even.

Case 1: m is odd. In this event, m^2 is odd. Since $2n^2$ is even for every n, there does not exist an integer n with $m^2 = 2n^2$.

Case 2: n is odd. Then $2n^2$ is an even integer that is not divisible by 4. By Lemma 1.4, $2n^2$ is not a perfect square.

This shows that if m and n are positive integers, then $m^2 \neq 2n^2$, which completes the proof. □

Equivalent Forms of Implication

Perhaps the most visual way to understand the conditional statement "If H, then C" is in terms of sets and subsets. Let $\mathcal{H}$ denote the set of all objects satisfying the hypothesis H and let $\mathcal{C}$ denote the set of all objects satisfying the conclusion C. The logical implication "If H, then C" takes the form $\mathcal{H} \subset \mathcal{C}$, see Figure 1.3 below. In words, "Every object that satisfies the hypothesis H also satisfies the conclusion C."

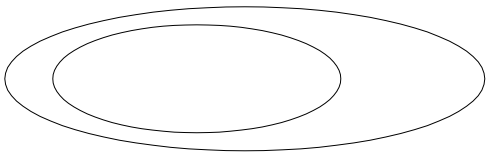


Figure 1.3: The Venn Diagram for "If H, then C".

Let $\mathcal{H}^c$ denote the set of objects that *do not* satisfy the hypothesis H, and let $\mathcal{C}^c$ denote the set of objects that *do not* satisfy the conclusion C. Thus $\mathcal{H}^c$ is the set-theoretic *complement* of the set $\mathcal{H}$. It should be clear from Figures 1.3 and 1.4 that $\mathcal{H} \subset \mathcal{C}$ means the same thing as $\mathcal{C}^c \subset \mathcal{H}^c$. The corresponding logical implication, "If not C, then not H," is called the *contrapositive* of the statement "If H, then C." Every implication is logically equivalent to its contrapositive. The contrapositive is sometimes stated "H only if C."

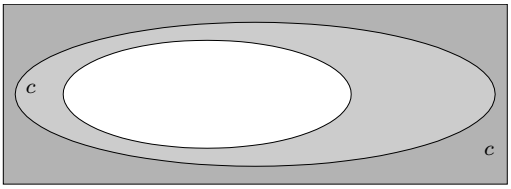


Figure 1.4: The contrapositive: "If not C, then not H".

It is common to use the notation $H \implies C$, read "H *implies* C," instead of the equivalent forms $\mathcal{H} \subset \mathcal{C}$ or "If H, then C." Logicians write $\neg C$ instead of "not C." In this notation, the contrapositive is written $\neg C \implies \neg H$.

Yet a third reformulation of $\mathcal{H} \subset \mathcal{C}$ is $\mathcal{H} \cap \mathcal{C}^c = \varnothing$. In words, there is no object that both satisfies the hypothesis H and *fails* to satisfy the conclusion C.

Statement	Name	Set Interpretation
If H, then C.	Direct Implication	$\mathcal{H} \subset \mathcal{C}$
If not C, then not H.	Contrapositive	$\mathcal{C}^c \subset \mathcal{H}^c$

Table 1.1: Direct implication and contrapositive.

The following sentences (each logically equivalent to $H \implies C$) are used interchangeably, usage being dictated primarily by style: If H, then C; C if H; H only if C; H is *sufficient* for C; C is *necessary* for H. Mathematical reading demands a great deal of attention to precise wording!

Example 1.5 The (valid) implication "If m and n are even integers, then $m + n$ is even" is written equivalently as

- $m + n$ is even if m and n are even.

- m and n are even only if $m+n$ is even.
- In order that $m+n$ be even, it suffices that m and n be even.
- In order that m and n be even, it is necessary that $m+n$ be even.
- If $m+n$ is not even, then m and n are *not both* even; in other words, *at least one* of m and n is not even.

Observe carefully that nothing in the last phrasing excludes the possibility that neither m nor n is even. The hypothesis is "$m+n$ is odd," and the conclusion is "m and n are not both even" (or "at least one of them is odd"). While it is *true* that both numbers cannot be odd (if their sum is not even), this fact is not asserted. You might say a true mathematical assertion *need not tell the whole truth.* □

Converse and Inverse

There are two other statements that are similar in appearance—*but not logically equivalent*—to $H \implies C$. The first, called the *converse*, is $C \implies H$; this asserts that if the conclusion is satisfied, then so is the hypothesis. The second, called the *inverse* $\neg H \implies \neg C$, asserts that if the hypothesis is not satisfied, then neither is the conclusion.

Statement	Name	Set Interpretation
If C, then H.	Converse	$\mathcal{C} \subset \mathcal{H}$
If not H, then not C.	Inverse	$\mathcal{H}^c \subset \mathcal{C}^c$

Table 1.2: Converse and inverse.

A very common mistake is to confuse a statement with its converse or inverse. Generally, a statement is not logically equivalent to its converse. In Example 1.5, the (non-valid) converse implication reads "If $m+n$ is even, then m and n are even." The incorrectness of the converse is exhibited by the existence of *counterexamples*: Indeed, the sum of two odd integers is even.

Methods of Proof and Disproof

Every theorem in mathematics is (equivalent to) one or more logical implications, though sometimes the logical implication is disguised by

the wording of the theorem. An "if and only if" (sometimes written "iff" or "$\Leftrightarrow$") statement is a pair of logical implications where each statement is the converse of the other. Loosely, a true iff statement *is the whole truth*. It is a stylistic tradition that in a definition, the word "if" is always taken to mean "iff." Thus Definition 1.1 above really means "An integer n is *even* if, and only if, there is an integer m such that $n = 2m$."

A *proof* is an argument used to establish the truth of a logical implication $H \implies C$. There are three methods of proof, corresponding to the three set-theoretic interpretations of the implication.

Direct Proof With the direct method, $\mathcal{H} \subset \mathcal{C}$ is proven by showing that if x is in $\mathcal{H}$, then x is in $\mathcal{C}$, or that "$x \in \mathcal{C}$ for every $x \in \mathcal{H}$." In words, choose a "generic" object x that satisfies the hypothesis H and show that this object must also satisfy the conclusion C.

Contraposition Proof by contraposition relies on the equivalence of $\mathcal{H} \subset \mathcal{C}$ and $\mathcal{C}^c \subset \mathcal{H}^c$. This method may be regarded either as a different method of proof, or as a direct proof of a different (but logically equivalent) statement. In this method, choose a "generic" object that fails to satisfy the conclusion and show that it fails to satisfy the hypothesis as well.

Contradiction Proof by contradiction relies on the equivalence of $\mathcal{H} \subset \mathcal{C}$ and $\mathcal{H} \cap \mathcal{C}^c = \varnothing$. The approach is to show that if some object simultaneously satisfies the hypothesis and fails to satisfy the conclusion, then mathematics is logically inconsistent: There is a statement P such that P and $\neg P$ are both true.

Example 1.6 The standard proof that there is no rational square root of 2 is by contradiction, and relies more heavily on Lemma 1.4. Assume m/n is in lowest terms, and that $(m/n)^2 = 2$, or $m^2 = 2n^2$. This equation implies that the even integer $2n^2$ is a perfect square, hence is a multiple of 4 by Lemma 1.4. Thus $m = 2\ell$, that is, m is even. Dividing $2n^2 = m^2 = 4\ell^2$ by 2 gives $n^2 = 2\ell^2$. Applying Lemma 1.4 again, we find that n^2 is a multiple of 4, so n is even. Thus m/n is not in lowest terms, contradicting the original assumption.

In summary, the argument above shows that if m/n is in lowest terms and $(m/n)^2 = 2$, then m/n is *not* in lowest terms. This shows

$(m/n)^2 = 2$ is impossible, that is, there is no rational square root of 2. □

Proof by contradiction tends to be awkward logically, and does not generally build a logical link between hypothesis and conclusion. For this reason, proof by contradiction should be regarded as a last resort. Happily, most proofs by contradiction can easily be re-written as proofs by contraposition, though as in Theorem 1.3 it may be necessary to reformulate the implication appropriately.

It is a common mistake, especially under exam pressure, to start a proof by assuming the conclusion. This amounts to assuming what is to be proved, and is clearly wrong. To emphasize:

> When proving a logical implication, the conclusion is *never assumed*.

Let us return for a moment to Theorem 1.2, which asserts (loosely) the impossibility of squaring the circle in Euclidean geometry. Here are the basic ideas of the proof: First it is shown that if a segment of length π could be constructed in finitely many steps with a straightedge and compass, then the real number π would satisfy a polynomial relation with rational coefficients—something like $\pi^2 - 10 = 0$ or $1 - \pi^2/6 + \pi^4/120 = 0$ (neither of these is correct). But for purely analytic reasons (that are, unfortunately, beyond the scope of this book), such a relation would imply existence of an integer between 0 and 1. Since no such integer exists, the purported construction is *impossible*, in the sense of being incompatible with basic properties of numbers.

Counterexamples The previous items deal with establishing the truth of a logical implication. The dual task, proving the falsity of a logical implication, is accomplished by means of *counterexamples*. A counterexample to the assertion $H \implies C$ is an object x that both satisfies the hypothesis and fails to satisfy the conclusion. Existence of such an x proves that the intersection $\mathcal{H} \cap \mathcal{C}^c$ is non-empty, so that $\mathcal{H} \subset \mathcal{C}$ is false, see Figure 1.5.

While the falsity of the assertion $H \implies C$ can be proven by finding an object that both satisfies the hypothesis and fails to satisfy the conclusion, the statement $H \implies C$ cannot be proven by finding an example y satisfying both the hypothesis and the conclusion; in Figure 1.6, such an example exists but the statement "$H \implies C$" is false. To prove a logical implication, it must be shown that *every* object satisfying the hypothesis satisfies the conclusion.

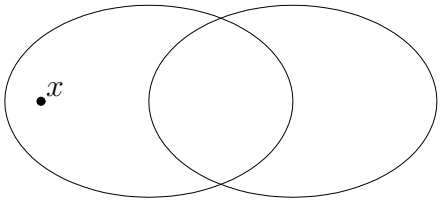


Figure 1.5: A Counterexample.

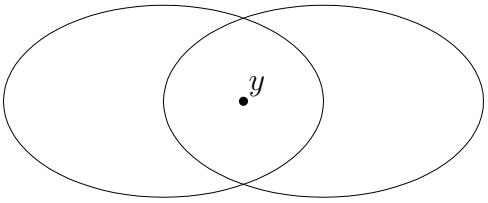


Figure 1.6: A Misleading Example.

Summary

Three themes run through all bodies of well-developed mathematical knowledge: Definitions, theorems, and examples. Definitions can be made freely, but *good* definitions are usually hard-won. When you encounter a definition for the first time, you should ask yourself: What are some examples and non-examples? What kinds of objects are distinguished by the definition? It is usually a bad definition that specifies only objects that can be characterized by a simpler alternative. It is also possible for a definition to compress an astounding amount of subtle information into a few deceptively simple criteria. The definition of the natural numbers—the first precise definition in this book—is an excellent example. Roughly, the quality of a definition is measured by its simplicity (which is connected to its ease of use) and the number of unexpected consequences it has.

Examples bring a definition to life; the worst definitions are those that have no examples (or no non-examples) because the definition is either logically inconsistent or tautological. Examples make definitions useful for modeling real-life situations, and definitions are usually chosen to make some real-world intuition precise. As noted earlier, it may be very difficult to verify whether or not a specific object is an example of a definition! For this reason, simple definitions are preferable: They are easier to verify when treating specific examples. However the advice of A. Einstein is germane: Make things as simple as possible, but no simpler.

Theorems are non-obvious consequences of definitions. They are useful for classifying examples (perhaps by reformulating a definition in an equivalent, non-obvious way), organizing logical relationships among concepts, and extending knowledge about classes of objects. Often knowledge about a particular object is gleaned by verifying that the object is an example of some definition, then using a theorem that guarantees all such objects have the desired property.

Mathematics is participatory, so these remarks may not mean much at this stage (unless they're intuitively obvious). If you return to them periodically throughout the book, you should find their meaning becoming clearer.

1.4 Calculus and the "Real World"

Calculus is often called the mathematical study of rates of change. The questions it addresses include finding the area enclosed by curves in a plane (and defining what is meant by "area" for a curved region), finding lines that are tangent to curves (and defining what is meant by "tangency"), calculating the energy needed to move an object in a non-constant gravitational field, predicting the rate of progression of a chemical reaction as the reactants are used up by the reaction, finding the speed of a stone dropped from a cliff (and defining what is meant by "speed at an instant of time"), and so forth. As a tool of the sciences, calculus comprises three majors facets:

- The "practical" side, which makes connection to the domains of physics, chemistry, biology, economics—the "real world"—and through which calculus acquires its largest group of consumers: scientists and engineers.

  Applications take the form of "Laws of Nature," which are equations whose solutions model some aspect of the physical world. In scientific modeling, there is always a trade-off between simplicity and accuracy. Newton's law of gravitation is simple, and adequate for most practical purposes, but it fails to explain some observable phenomena. Einstein's general theory of relativity refines and extends Newton's law, but it requires difficult and subtle mathematics in order to perform predictive calculations.

- The "calculational" aspect, which allows intuitive ideas about "infinitely small" quantities, "points" in space, and "instants" of time, to be converted into symbolic expressions and manipulated to obtain useful answers to the types of questions described above.

- The "theoretical" foundation, which defines differentials and integrals in terms of set theory (the "machine language" of mathematics) and proves that the manipulations are logically consistent, so

that the answers obtained are in some sense reasonable even if the technical details and intuition do not coincide exactly.

Each of these aspects is important in its own right, and each supports the others, though most users do not need to understand the foundations in order to use the applications. (To use a mechanical metaphor, you can drive a car without being able to rebuild the engine.) However, it is better to have true understanding of a subject rather than mere familiarity; true knowledge is flexible, and can be applied correctly in novel situations, while familiarity is limited to situations encountered previously.

Modeling Physics with Mathematics

A simple physical example will serve to illustrate these facets as well as several related issues that arise. Suppose a stone is dropped from a cliff y_0 meters high; how many seconds pass before the stone hits the ground, and how fast will the stone be moving at impact? In elementary physics, these answers are encoded in a formula for the height $y(t)$ of the stone after t seconds have passed:

$$y(t) = y_0 - 4.9t^2. \tag{1.1}$$

It is important to remember that this model is not an exact description of reality. We next discuss, in some detail, the assumptions that go into this model, to illustrate the practical way mathematics interfaces with the sciences.

In order to use mathematics to describe natural phenomena, a back-and-forth procedure of approximation and mathematical modeling must be employed. In a situation as simple as that of a falling stone, these mental machinations are often made unconsciously, but in complicated and novel problems it is necessary to understand how to translate from mathematics to "the real world" and back.

As a first approximation, the stone will be regarded as a *point particle*, imbued with no attributes other than position and mass, and the earth's surface will be modeled by a flat plane. Barring effects of wind and Coreolis deflection, the stone drops vertically, so the motion of the stone is determined by knowing, for each time t (measured in seconds after its release, for example), its height $y(t)$ (measured in meters, say) above the ground. Even in informal English the height of the stone is said to be "a function of time." The mathematical concept of a *function*

is fundamental, and is discussed at length in Chapter 3. Because the quantity of interest—the height of the stone at time t—is determined by a single number, this model is said to be a "one-variable" problem.

To say anything further, it is necessary to borrow some results from physics; mathematics says absolutely nothing about the way stones fall, nor in general about anything other than mathematics. In our idealized situation, the motion of the stone is governed by Newton's laws of motion: there are "forces" acting (gravitation and air resistance are the most important ones), and these determine the "acceleration" of the stone. Acceleration is a concept of differential calculus: the *velocity* of the stone is its rate of change of position (in units of meters per second), while the *acceleration* is the rate of change of the velocity (in " 'meters per second' per second"). In the Newtonian model, the forces acting on the stone determine its behavior, and it is this predictive power that answers the original questions.

It is convenient to make a couple of idealizations:

- The acceleration due to gravity is constant during the stone's fall.
- There is no air resistance.

The first assumption is justified as follows. According to Newton's law of gravitation, the force acting on the stone is a certain constant G times the mass m of the stone times the mass M of the earth, divided by the square of the distance $R(t)$ from the stone to the center of the earth at time t. According to Newton's third law of motion, the net force on the stone is equal to the mass of the stone times its acceleration. In symbols,

$$F = -G\frac{mM}{R(t)^2} = ma; \tag{1.2}$$

the minus sign indicates that the force is directed toward the center of the earth. Because the distance the stone falls is very small compared to the radius R of the earth, the ratio $R(t)/R$ is very nearly equal to 1 throughout the stone's fall, so the denominator in equation (1.2) may be replaced by R^2 without much loss of accuracy. The assumption that there is no air resistance is not realistic, but modeling the air resistance on a solid body is horrendously complicated even if the body is perfectly spherical (another unrealistic assumption). However, the point to be made concerns modeling, and neglecting air resistance illustrates this

point nicely: Sometimes you must make unrealistic simplifications to get a tractable calculation.

By equation (1.2), the acceleration of the stone is (approximately!) $a = -GM/R^2$. This number has been *measured* to be roughly -9.8 meters per second per second at the surface of the earth. At last we are ready to use calculus. The technique to be used is called "integration" and is studied at length starting in Chapter 7. Here we give the intuitive description; nothing in the next paragraphs should be taken too literally.

The acceleration of the stone—the rate of change of velocity—is constant, so in each instant of time the velocity increases by the same amount, and after t seconds is equal to $v(t) = v_0 - at$ meters per second; v_0 is the initial velocity, which is zero because the stone was dropped.

Now, in the same instant of time, the stone falls a distance of $-at\,dt$ meters. Adding up these infinitesimal distances (and omitting the details) leads at last to (1.1). To repeat, the height of the stone before impact is

$$y(t) = y_0 - \frac{1}{2}at^2 = y_0 - 4.9\,t^2.$$

As a sanity check, let us verify the correctness of this formula: In the instant of time between t and $t + dt$, the stone falls from height $y(t)$ to $y(t + dt)$, a distance of

$$dy := y(t + dt) - y(t) = -\frac{1}{2}a\Big((t + dt)^2 - t^2\Big) = -\frac{1}{2}a(2t\,dt + dt^2).$$

Dividing by dt gives the stone an instantaneous velocity of

$$\frac{dy}{dt} = -at - \frac{a}{2}dt, \tag{1.3}$$

which is essentially $v(t) = -at$ because dt is vanishingly small.

To emphasize once more, equation (1.1) is a *model* for the height of a dropped stone t seconds after its release from a height of y_0 meters. Armed with this formula, we can answer the question, "When does the stone land?" because the impact of the stone corresponds to the condition $y(t) = 0$ in the model and this equation is easily solved for t in terms of the initial height y_0. The stone's impact speed is even easier to find: at_0 meters per second (the discarded minus sign merely indicates that the stone hits the ground while moving downward). In terms of the initial height y_0 (in meters),

$$\text{Impact time} = \sqrt{\frac{y_0}{4.9}}, \qquad \text{Impact velocity} = 2\sqrt{4.9y_0}.$$

For most calculus courses this is the end of the story, and indeed, it is hoped that none of the mathematics or physics was unfamiliar. The point in going through this simple example in such detail was to point out the features of modeling a real-world situation, and to mention the interesting and controversial steps that occurred in obtaining the velocity of the stone from equation (1.2) and verifying that equation (1.1) really does lead to the computed velocity.

Velocity is supposed to represent "instantaneous rate of change of position," but what exactly does this mean? (If this question cannot be answered, then there is no reason to believe the equation $v = -at$ is meaningful or useful!) The ancient Greek Zeno of Elea discovered the following apparent paradox. Imagine the falling stone at "an instant of time." The stone has a definite location, and is indistinguishable from a stationary stone at the same height. More concretely, imagine an infinitely fast camera that captures literal instants of time. Then there is no way to distinguish a moving object from a stationary object on the basis of a single photograph. But since this argument can be made at every instant of time, there is no difference between moving and standing still! Yet the falling stone *does* fall rather than levitating, and motion is patently possible. Where did the argument go astray? To see intuitively why there is no paradox, imagine a *very fast* camera that can capture arbitrarily small intervals of time; an exposure of a thousandth of a second, a billionth of a second, or a billionth of a billionth of a billionth of a second is possible. To such a camera, a moving stone and a stationary stone *do not* appear identical; the moving stone makes a slightly blurred image because its position changes while the shutter is open. Of course, the image of the falling stone becomes sharper as the shutter speed is increased, but the resulting picture is *never identical* to a photograph of a stationary stone. Additionally, the distance traveled by the stone *divided by the exposure time* gets "closer and closer to a limiting value" as the shutter speed increases; this "limiting value" has units of meters per second, and is interpreted as the "instantaneous velocity" of the stone. Effectively, this limiting procedure "magnifies the time scale by a factor of ∞." This explanation is substantially incomplete, because the phrases in quotes have not been defined. Intuitively, the points are that:

- States of rest and motion can be distinguished over arbitrarily small intervals of time, even though they cannot be distinguished at a single instant;

- If we look at smaller and smaller intervals of time, motion looks more and more as if it is at constant speed; geometrically, the graph of position as a function of time looks more and more like a straight line as we "zoom in" at time t.

Though these remarks are imprecise, they contain the germ of precise definitions in which a logically consistent mathematical theory of rates of change can be formulated. The naive idea of "instantaneous velocity" (distance divided by time) involves the meaningless expression $0/0$. The concept of "limits" (Chapter 4) neatly circumvents this difficulty and permits mathematical treatment of instantaneous speed.

Several other issues have also arisen implicitly, though they are mostly non-mathematical. "Instants of time" are physically meaningless, as are spatial points, though both concepts serve as useful idealizations that are quite accurate for macroscopic events. In making mathematical models of "real" situations, it is always necessary to neglect certain effects, to forget some attributes of the objects under consideration, and even to make purely mathematical approximations (for example, in solving numerical models with a digital computer). Mathematics is a precise and detailed language that happens to be useful for describing many observed phenomena. Many of the laws of nature can be expressed conveniently in the language of *differential equations*, which relate quantities and their rates of change. Sciences such as physics and chemistry attempt to relate outcomes of experiments with mathematical descriptions so that the results can be predicted in advance or otherwise understood. These mathematical models—so-called Laws of Nature—may be quite accurate, but none as yet is all-encompassing or completely accurate to the limits of measurement.

There is good reason to assert that mathematics is not "real" in a physical sense; it is a tool or language that our minds use to construct accurate, predictive models of reality. The aim of this book is to show how the mathematics underlying calculus is logically consistent by building it from set theory, while giving interesting and substantial applications of these powerful mathematical techniques.

Exercises

Some of these exercises are fairly traditional, and assume you are familiar with standard mathematical notation. Others are designed to make you think about language, mental models, and semantics. The

only way to learn to "speak Mathematics" is through practice; writing, reformulating, and thinking. Familiarity can be acquired through reading, but originality can only come through participation. Additionally, Mathematics uses English words and grammar (in this book, at least), but is not English. A few of the exercises illustrate important differences. Mathematical definitions and theorems tend to be stated in a "logical" form that our brains are not normally adept at understanding, see Exercise 1.10. For most students, it is a psychological hurdle to recognize that Mathematics is a language with a fairly rigid syntax, that colloquial speech is often substantially imprecise, and that something suggested by association is not linked by logic, see Exercises 1.3, 1.4, and 1.9. (Almost all advertising relies on the confusion of association with a logical link.)

Exercise 1.1 Prove that if a whole number k is a perfect square, then either k or $k-1$ is divisible by 4. (Compare Lemma 1.4.) ◇

Exercise 1.2 Let X and Y be sets. The *symmetric difference* $X \triangle Y$ is defined to be $(X \setminus Y) \cup (Y \setminus X)$. Illustrate the definition with a Venn diagram; prove that

$$X \triangle Y = (X \cup Y) \setminus (X \cap Y),$$

and that the symmetric difference corresponds to the Boolean operation xor ("exclusive or").

Note: To prove two sets X and Y are equal, you must show two things: $X \subset Y$, and $Y \subset X$. ◇

Exercise 1.3 Someone holding a bag of blue marbles says, "Every marble in this bag is either red or blue." Is this a true statement? If not, why not? If so, does it "tell the whole truth"? Explain. ◇

Exercise 1.4 R. M. Waldo, the tallest documented human, was just under 9 feet in height. Assume for this question that he was the tallest human ever to live, and that he was exactly 9 feet tall.

(a) Consider the claim, "Humans are at most 12 feet tall." Is this claim true? If not, why not? If so, can you write it as an "If..., then..." statement?

(b) Consider the claim, "Humans are at most 9 feet tall." Does this claim "tell the whole truth"? If so, in precisely what sense does it do so?

(c) Does the assertion in part (a) imply the assertion in part (b), or *vice versa*?

Telling the truth mathematically is not the same as telling the truth in a Court of Law. (Standards of proof are different as well, but that is another issue.) ◇

Exercise 1.5 Though it seems strange at first, a set can be be an element of itself. One early problem with set theory was *Russell's paradox*: Let X be the set of all sets that are not elements of themselves. Prove that X is an element of itself iff it is *not* an element of itself.

This problem was fixed by exercising more care with the notion of a "set": There must be a universe fixed at the outset, and the "set of all sets" is not a "set" but a larger object called a "class." ◇

Exercise 1.6 Some of the greatest achievements of 20th Century logic were made by K. Gödel. Among other things, Gödel formulated a statement in arithmetic that could be *interpreted* as saying, "This statement cannot be proven." Assuming there is no contradiction in mathematics, prove that Gödel's statement is *true* but *unprovable* (in the axiomatic framework in which it is formulated). This demolished the cherished idea that every mathematical truth can be proven in some fixed axiomatic system. ◇

Many statements in analysis involve the *quantifiers* "for every" (the *universal quantifier* $\forall$) and "there exists" (the *existential quantifier* $\exists$). We will not use these symbols, though you are cautioned that some people are fond of writing, e.g.,

$$(\forall \varepsilon > 0)\ (\exists \delta > 0)\ |x - x_0| < \delta \implies |f(x) - f(x_0)| < \varepsilon.$$

The next exercise introduces some basic properties of quantifiers.

Exercise 1.7 Consider the quantified sentences:

(a) For every marble x in the bag, x is blue.

(b) There exists a red marble y in the bag.

(c) For every real $x > 0$, there exists a natural number n with $1/n < x$.

Give the negation of each sentence. Express (a) and (c) as conditional statements, and give their contrapositives. Express the negations of (b) and (c) as conditional statements, and give their contrapositives.

Notice that "for every" and "there exists" are exchanged under negation. ◇

Exercise 1.8 A game-show host presents the contestant with the equation "$a^2 + b^2 = c^2$." The contestant replies, "What is the Pythagorean Theorem?"

(a) Is this really the correct question? If not, can you append a clause to give a question that would satisfy a mathematician?

(b) State the Pythagorean Theorem in if-then form.

A theorem is not merely its conclusion. Despite this, after A. Wiles announced the proof of Fermat's Last Theorem, the Mathematical Sciences Research Institute (MSRI) in Berkeley produced T-shirts bearing the message, "$a^n + b^n = c^n$: NOT!" ◇

Exercise 1.9 The President, a law-abiding citizen who always tells the truth, has time for one more Yes/No question at a press conference. In an attempt to embarrass the President, a reporter asks, "Have you stopped offering illegal drugs to visiting Heads of State?"

(a) Which answer ("Yes" or "No") is logically truthful?

(b) Suppose the President answers "Yes". Can the public conclude that the President has offered illegal drugs to visiting Heads of State? What if the answer is "No"?

(c) Explain why both answers are embarrassing.

If the President were a Zen Buddhist, she might reply, "mu,"[5] meaning "Your question is too flawed in its hypotheses to answer meaningfully." ◇

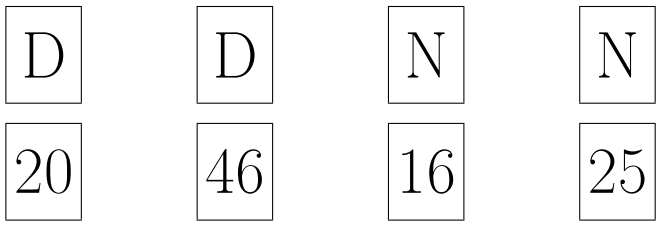
Exercise 1.10 One striking peculiarity of the human brain is that it is "better" at seeing certain situations in an emotional light than it is at understanding an equivalent logical formulation. Here is an example.

(a) Each card in a deck is printed with a letter on one side and a number on the other. Precisely, the letter is either "D" or "N" and the number is a whole number between 16 and 70 inclusive. There is no restriction on the combinations in which cards are printed. Your job is to assess whether or not cards satisfy the

[5]Pronounced "moo"

single criterion: "Every 'D' card has a number greater than or equal to 21 printed on the reverse." You are also to separate cards that satisfy this criterion from those that do not.

Write the criterion as an "If..., then..." statement, and determine which of the following cards satisfy the criterion:



(i) (ii) (iii) (iv)

(b) You are shown four cards:

(i) (ii) (iii) (iv)

Which cards must be turned over to determine whether or not they satisfy the criterion of part (a)?

(c) The legal drinking age in a certain state is 21. Your job at a gathering is to ensure that no one under 21 years of age is drinking alcohol, and to report those that are. A group of four people consists of a 20 year old who is drinking, a 46 year old who is drinking, a 16 year old who is not drinking, and a 25 year old who is not drinking. Which of these people is/are violating the law?

After reporting this incident, you find four people at the bar: An 18 year old and a 35 year old with their backs to you, and two people of unknown age, one of whom is drinking. From which people do you need further information to see whether or not they are violating the law?

(d) Explain why the card question is logically equivalent to the drinking question.

Which did you find easier to answer correctly? ◇

Chapter 2

Numbers

There is a pervasive but incorrect perception that mathematics is the study of numbers; some mathematicians even joke that the public believes mathematical research consists of extending multiplication tables to higher and higher factors. This misapprehension is fostered by school courses that treat routine calculation as the primary goal of mathematics. In fact, calculation is a *skill*, whose relationship to mathematics analogous to the relationship of spelling to literary composition.

In school, you learned about various kinds of numbers: the counting (*natural*) numbers, whole numbers (*integers*), fractions (*rational numbers*), and decimals (*real numbers*). You may even have been introduced to *complex numbers*, or at least to their most famous non-real member, i, a square root of -1. One goal of this chapter is to (re-)acquaint you with these sets of numbers, and to present their abstract properties—the associative, commutative, and distributive laws of arithmetic, properties of inequalities, and so forth. At the same time, you will see (in outline) how these sets of numbers are constructed from set theory, and discover that the way you have learned about numbers so far is almost purely *notational*. Nothing about the integers requires base 10 notation, and nothing about the real numbers forces us to use infinite decimals to represent them. You will learn to view numbers *notionally*, in terms of axioms that abstract their properties. The philosophical question "What *is* a real number?" will evaporate, leaving behind the answer "An element of a set that obeys several axioms."

One misnomer should be dispelled immediately: Though $\sqrt{2}$ is a "real" number while $\sqrt{-1}$ is an "imaginary" number, each of these symbols represents a mathematical abstraction, and neither has an existence more or less "real" than the other. No *physical* quantity can

definitively represent $\sqrt{2}$, because measurements cannot be made with arbitrary accuracy. If you take the square root of 2 on a calculator, you get a *rational* number whose square is often noticeably not equal to 2. It cannot even be argued that $\sqrt{2}$ has a geometric meaning (the diagonal of a unit square) and $\sqrt{-1}$ doesn't; the imaginary unit can be viewed perfectly well as a 1/4-rotation of a plane. Performing such a rotation twice (squaring) reflects the plane through its origin, which is tantamount to multiplication by -1. The geometric picture of complex multiplication is "real" enough that it can be used to interpret Euler's famous equation $e^{i\pi} + 1 = 0$, as we shall see in Chapter 15.

The natural numbers are simple enough to be defined directly in terms of sets, but more complicated number systems—the integers, rationals, reals, and complex numbers—are defined successively, in terms of the previous type of numbers. There is a fringe benefit: The techniques of recursive definition, mathematical induction, and equivalence classes, which arise naturally in constructing the integers from set theory, are important and useful throughout mathematics. By far the most complicated step is the definition of real numbers in terms of rational numbers. If the recursive definitions are expanded, a *single* real number is a mind-bogglingly complicated set. Luckily, it is never necessary to work with "expanded" definitions; the abstract properties satisfied by the set of real numbers are perfectly adequate.

To make a computer analogy, sets are bits, natural numbers are bytes, the integers and rational numbers are words, the real and complex numbers are assembly language, and calculus itself is an application program. At each stage, the objects of interest are built from the objects one level down. No sane person would write a spreadsheet program in assembly language, and no sane person would attempt to interpret an integral in terms of sets at the level of natural numbers. The recursive nature of calculus (or programming) allows you to forget about the details of implementation, and concentrate on the properties your building blocks possess.

2.1 Natural Numbers

L. Kroeneker, the 19th Century mathematician, said (in free translation) that "the natural numbers alone were created by God; all others are the work of Man [sic]." A more modern (and secular) phrasing is mathematics forces us to study the counting numbers, but real numbers

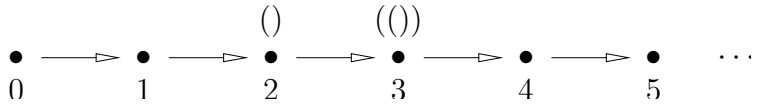


Figure 2.1: The set of natural numbers.

are a human invention.

You have met the set $\mathbf{N}$ of natural numbers as the set of counting numbers, starting with 0. Abstractly, N is "an infinite list" characterized by three properties:

N.1 There exists an *initial element*, denoted for the moment by ϕ.

N.2 There is a notion of *successorship*: Every natural number n has a unique "successor" $\star n$, and every natural number other than ϕ is the successor of a unique natural number, its *predecessor*.

N.3 For every non-empty subset $A \subset \mathbf{N}$, there exists a *smallest element*, namely an element $n_0 \in A$ such that every other element of A arises in the chain of successorship beginning with n_0.

The set $\mathbf{N}$ is depicted in Figure 2.1: Successorship is denoted by arrows, and elements of $\mathbf{N}$ are denoted both in the usual way (Hindu-Arabic numerals) and by using the notion of successors. The ellipsis indicates that "the pattern continues forever".

Property **N**.1 says $\mathbf{N}$ is non-empty, while **N**.3, the *well-ordering property*, is the basis for the method of mathematical induction. Conceivably, **N**.1–**N**.3 are logically inconsistent. To show that they are not, we will construct a set $\mathbf{N}$ and a notion of successorship that satisfy **N**.1–**N**.3. For the most part, we regard **N**.1–**N**.3 as *axioms*—statements whose truth is unquestioned. In other words, we will "forget the details of the implementation" of Theorem 2.1 and take **N**.1–**N**.3 as the starting point for deducing properties of the natural numbers.

Theorem 2.1. *There exists a set* $\mathbf{N}$*, with a notion of successorship, that satisfies Properties* **N**.1–**N**.3*. This set is unique up to an order-preserving isomorphism.*

The term "order-preserving isomorphism" is explained in Chapter 3, but in the present situation means "a relabelling of elements that preserves the notion of successorship".

Proof. (Sketch) There is a unique set $\varnothing$ that contains no elements. *Define* ϕ to be the set $\varnothing$. Now define successorship: If $n \in \mathbf{N}$ has been defined (as a set), then the *successor* of n is defined to be the set

$$\star n := n \cup \{n\}. \tag{2.1}$$

Unwrapping the definition gives, in Hindu-Arabic notation,

$$\begin{aligned} 1 &= \{\varnothing\} \\ 2 &= 1 \cup \{1\} = \big\{\varnothing,\ \{\varnothing\}\big\} \\ 3 &= 2 \cup \{2\} = \big\{\varnothing,\ \{\varnothing\},\ \big\{\varnothing,\ \{\varnothing\}\big\}\ \big\} \\ &\ \ \vdots \end{aligned}$$

Note that $n \subset \star n$ as sets, and that (for example) the set "3" has three elements: $\varnothing$, $\{\varnothing\}$, and $\big\{\varnothing,\ \{\varnothing\}\big\}$.

In a careful proof, there are many details to check, such as construction of a universal set in which the construction takes place. One must also take care to use only axioms of set theory, and not to rely on (possibly wishful!) intuition. However, the following assertions should be clear:

- For each pair n and m of distinct natural numbers (regarded as sets), either $n \subset m$ and m arises in the chain of successorship beginning with n, or *vice versa*. The usual meaning of "less than" for natural numbers is exactly proper inclusion of sets.

- The "number of elements" of the set n is exactly what is normally meant by the natural number n. This is because the set $\star n$ contains exactly one more element than the *set* n, namely the *element* n. (Do not confuse elements, nested within one sets of braces, with the number of "$\varnothing$" symbols.)

Property **N**.1 is true by construction, as is most of **N**.2; the only non-obvious detail is that a natural number cannot have two different predecessors. But granting the facts just asserted, if $\star n_1 = \star n_2$, then (relabelling if necessary) $n_1 \subset n_2$. If n_1 and n_2 were distinct, then we would have $\star n_1 \subseteq n_2$, which is impossible because $\star n_1 = \star n_2$.

To see that every non-empty set $A \subset \mathbf{N}$ has s smallest element, start at ϕ and take successors. At each stage, you are either in A or not. If you start in A, there is nothing to prove, while if you never arrive at an element of A, then by construction of $\mathbf{N}$ the set A is

empty. Otherwise, there is a first time that successorship yields an element n_0 of A, and every other element of A must arise in the chain of successorship starting with n_0. □

Though the recursive structure of (2.1) is simple, the expanded notation is unusable: Writing out the integer 100 is impossible because the number of "$\varnothing$" symbols *doubles* with each succession. "Hash marks", where for example 𝍸 denotes "5", represent natural numbers more efficiently (in a manner similar to Roman numerals), but are still inadequate for modern use. Hindu-Arabic numerals are extremely compact; each additional digit describes numbers *ten times as large*, and the use of positional context to ascribe value to a digit (the ones column, tens column, and so forth) facilitates calculation, making it easy to write, add, and multiply enormous natural numbers.

Recursive Definition, and Induction

Consider the problem of writing a computer program completely from scratch. Ordinarily, a programmer picks a "computer language" such as C, writes a formally-structured but human-readable "source code file", then uses a special program called a *compiler* to convert the source code into a pattern of bytes that a machine can execute. Now, a modern C compiler is an extremely complicated, sophisticated program, something that is too complex to write from scratch. How would someone get started without a compiler? The answer is that they would first write a small, not very featureful compiler in machine language, then *use it to compile a more powerful compiler*. Next they would take their "second-stage" compiler and write a full-blown C compiler. Thusly equipped, they would write the program they set out to create.

By analogy, suppose we wish to construct the set of real numbers from scratch. Our construction of $\mathbf{N}$ above is something like a bare computer, capable of being programmed but having no software at all. The set of natural numbers does not come equipped with the arithmetic operations of addition, multiplication, and exponentiation; these must be constructed from the notion of successorship, and are analogous to our hand-written, "first-stage" compiler. Armed with natural numbers and arithmetic operations, we proceed to construct the integers and the rational numbers, which are analogous to the successive compilers; only then are we ready to construct the set of real numbers.

The ideas introduced above embody *recursive definition*, in which

a sequence of objects or structures is defined, each in terms of the previous. Even the construction of arithmetic operations on $\mathbf{N}$ has a strongly recursive flavor. *Addition* is defined to be iterated succession: "$2 + 3$" means "the successor of the successor of the successor of 2" ('take the successor of 2' 3 times), see equation (2.2). Once addition is available, *multiplication* is defined to be iterated addition, and *exponentiation* is defined to be iterated multiplication. The familiar properties of arithmetic will be proven using *mathematical induction*, a technique for establishing the truth of infinitely many assertions that are suitably "logically linked". From these structures flows a vast and deep river of ideas and theorems, whose extent is by no means entirely mapped. Fermat's Last Theorem was established only in 1994; other assertions about $\mathbf{N}$, such as Goldbach's Conjecture[1] or the Twin Primes Conjecture[2], remain open.

Addition of Natural Numbers

The intuition behind addition of natural numbers is "agglomeration of heaps", as in $\bullet\bullet + \bullet\bullet\bullet = \bullet\bullet\bullet\bullet\bullet$. You should keep in mind that the following discussion is just a formalization of this idea, and remember the Goethe quote.

Suppose we wish to program a computer to add two natural numbers, using the definition of addition as iterated successorship. While intuitively the expression "$m + n$" means "start with m and take the successor n times," such a prescription is not immediately helpful, because a computer has a fixed processor stack, while the number n can be arbitrarily large. What is needed is a *procedure* that requires only a fixed amount of processor memory.

Imagine, for the moment, that you do not know the meaning of "$+$". Any use of the symbol must be explained in terms of natural numbers and successorship, and in particular there is no reason to assume formal properties like $m + n = n + m$. Let m be a natural number, and define $m + 1 = \star m$. Now, if n is a natural number, and if the expression "$m + n$" has been defined for all $m \in \mathbf{N}$, then we *define*

$$m + (n + 1) := (m + n) + 1 \qquad \text{for all } m \in \mathbf{N}. \tag{2.2}$$

It is necessary to make a separate definition for $m + 1$ to start the process. In less suggestive notation, $m + (\star n) := \star(m + n)$ for all

[1] Every even number greater than 4 is a sum of two odd primes.
[2] There exist infinitely many pairs of primes of the form $\{p,\ p + 2\}$.

$m \in \mathbf{N}$. It is not obvious that $m + n = n + m$ for all m and n; this is part of Theorem 2.6 below.

Equation (2.2) provides an *algorithm* for "taking the n-fold iterated successor": Begin with two "heaps of stones" and move stones from the second heap to the first until no stones are left in the second heap. In pseudocode,

```
Let m and n be natural numbers;
While ( n > 0 ) {
  Replace m by its successor;
  Replace n by its predecessor;
}
Return m;
```

Mathematical Induction

Theorem 2.2, the *Principle of Mathematical Induction*, is one of the most important places natural numbers arise in analysis. In many settings, an infinite string of assertions can be proven by establishing the truth of one of them (usually but not always the first one) *and* showing that each statement implies the next. Induction is particularly well-suited to situations in which something has been defined recursively.

Theorem 2.2. *Suppose that to each natural number $n \in \mathbf{N}$ is associated a sentence $P(n)$, and that*

I.1 (*Base case*) *The sentence $P(n_0)$ is true for some $n_0 \in \mathbf{N}$;*

I.2 (*Inductive Step*) *For each natural number $k \geq n_0$, the truth of $P(k)$ implies the truth of $P(k+1)$.*

Then each sentence $P(n)$ with $n \geq n_0$ is true.

Proof. Intuitively, $P(n_0)$ is true by hypothesis, so the inductive step says that $P(n_0+1)$ must also be true. But then the inductive step says that $P(n_0+2)$ is true, and so on *ad infinitum*. Thus all the subsequent sentences are true.

Formally, let $A \subset \mathbf{N}$ be the set of $n \geq n_0$ for which $P(n)$ is false. We wish to show that if the hypotheses of the theorem are satisfied, then A is empty. Equivalently, we may establish the contrapositive: If A is not empty, then the hypotheses of the theorem are not satisfied.

Assume A is not empty. By Property **N**.3, there is a smallest element of A. Let n be its predecessor. By definition of A, the statement $P(n)$ is true, while $P(n+1)$ is false; thus the inductive step fails for $k = n$, so the hypotheses of the theorem are not satisfied. □

Example 2.3 Suppose $(a_k) = a_0, a_1, a_2, \ldots$ is a sequence of numbers. We use recursive definition to define the sequence of *partial sums* as follows. First we set $S_0 = a_0$; then, for $n \geq 0$ we define $S_{n+1} = S_n + a_{n+1}$. This definition successively expands to $S_0 = a_0$, $S_1 = a_0 + a_1$, $S_2 = a_0 + a_1 + a_2$, and so on. *Summation notation* is extremely useful in this context; we write

$$S_n = a_0 + a_1 + \cdots + a_n = \sum_{k=0}^{n} a_k \text{ for } n \geq 0.$$

The expression on the right is read, "the sum from $k = 0$ to n of a_k", and is called the nth *partial sum* of the sequence (a_k). The recursive definition of the partial sums looks like

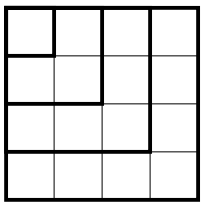
$$\sum_{k=0}^{n+1} a_k = a_{n+1} + \sum_{k=0}^{n} a_k \text{ for } n \geq 0. \tag{2.3}$$

The symbol k, called a *dummy index (of summation)*, is a "local variable" that has no meaning outside the summation sign. It is simply a placeholder to remind us that we are adding up a finite sequence of terms, and can be replaced by any letter not already in use, such as i or j. By contrast, n is related to the number of terms, and is not a local variable.

It is important to become fluent with summation notation. We will encounter plenty of examples in due time. Exercise 2.4 provides further practice. □

Example 2.4 Suppose we wish to find a formula for the sum of the first n odd integers. The first step is to guess the answer! Though mathematics is deductive, it is often *discovered* by trial and error or educated guessing. The role of proof is to verify that a guess is correct. The odd integers are 1, 3, 5, 7, and so forth; the nth odd integer is $2n-1$. The first few sums are

$$\begin{aligned} 1 &= 1 \\ 1 + 3 &= 4 \\ 1 + 3 + 5 &= 9 \\ 1 + 3 + 5 + 7 &= 16 \end{aligned}$$



On the basis of this evidence, it looks like the sum of the first n odd integers is n^2. However, despite the compelling picture, no amount of case-by-case checking will suffice to prove this claim for all $n \in \mathbf{N}$, because infinitely many claims are being made. To attempt a proof by induction, consider the sentence

$$P(n): \qquad \sum_{i=1}^{n}(2i-1) = 1+3+5+\cdots+(2n-1) = n^2.$$

At this stage, we do not know if some or all of the sentences $P(n)$ are true; we will attempt to demonstrate their truth by showing that conditions I.1 and I.2 hold.

To verify I.1, replace n by 1 everywhere in $P(n)$; this yields the sentence $1 = 1^2$, an obvious truth. To establish the induction step, assume the kth sentence $P(k)$ is true. In this example, assume that

$$P(k): \qquad \sum_{i=1}^{k}(2i-1) = k^2.$$

The left-hand side is the sum of the first k odd numbers; to get the sum of the first $(k+1)$ odd numbers, we add the $(k+1)$st odd number $2(k+1)-1 = 2k+1$ to both sides of the equation, and perform algebra:

$$\begin{aligned}\sum_{i=1}^{k+1}(2i-1) &= \Big(\sum_{i=1}^{k}(2i-1)\Big) + (2k+1)\\ &= k^2 + (2k+1) = (k+1)^2.\end{aligned}$$

The second equality uses the inductive hypothesis $P(k)$, while the last step is an algebraic identity. The entire equation is the assertion $P(k+1)$. Thus, if $P(k)$ is true, then $P(k+1)$ is also true. By the Induction Principle, $P(n)$ is true for all $n \in \mathbf{N}$.

The end result of this argument is a useful fact, and is said to express the given sum "in closed form." To find the sum of the first 1000 odd numbers, it is not necessary to perform the addition, but merely to square 1000. □

Mathematicians are famous for reducing a problem to one they have already solved. In principle, the more difficult problem is then also solved; in practice, a complete solution may be extremely complicated,

because the earlier problem may rely on solution of an even simpler problem, and so forth.

Example 2.5 The Tower of Hanoi puzzle consists of 7 disks of decreasing size that are initially in a stack and which can be placed on any of three spindles, Figure 2.2. The object is to move the stack of disks from one spot to another, subject to two rules:

(i) Only one disk may be transferred at a time.

(ii) A disk may rest only on a larger disk or on an empty spot.

The question is twofold: Determine how to transfer the disks, and find the smallest number of individual transfers that move the entire stack.

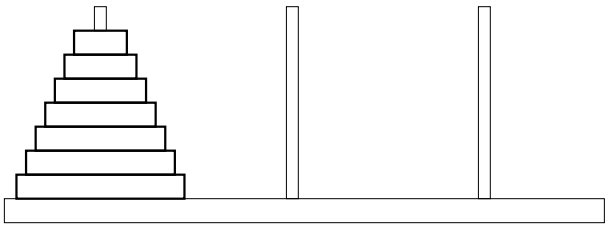


Figure 2.2: The Tower of Hanoi.

More generally, the game can be played with n disks. Before you read further, you should try to solve the puzzle for small values of n; coins of varying denominations make good disks. When $n = 1$ or $n = 2$, the solution is obvious, and for $n = 3$ or $n = 4$ the solution should be easy to find. According to legend, there is a Brahmin monastery where the monks toil ceaselessly to transfer a stack of 64 disks from one spot to another, subject to the rules above; when they complete their task, the universe will come to an end with a thunderclap. We need not fear the truth of this legend, as will soon become apparent.

The Tower of Hanoi has a beautiful and simple recursive structure. Let us take a managerial approach to the general problem: Suppose we knew how to move a stack of $(n-1)$ disks between any pair of spindles. We could then solve the problem by moving the top $(n-1)$ disks from spindle 1 to 2 (Figure 2.3; this requires many individual transfers, but may be regarded as a single operation), moving the largest disk from 1 to 3, and finally moving the stack from 2 to 3. This reduces solving the

n disk Tower of Hanoi to solving the $(n-1)$ disk tower. The 1 disk tower is trivial. If you know a programming language, you may enjoy implementing this recursive algorithm and seeing how long it takes to run with a modest number of disks.

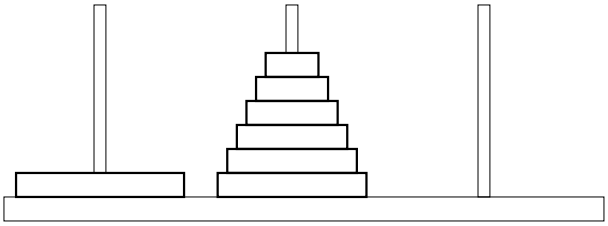


Figure 2.3: Solving the Tower of Hanoi puzzle recursively.

To see how many individual transfers must take place, examine the recursive structure of the solution more carefully: Imagine there are n people, labelled 1 through n, and that person j knows how to solve the j disk tower. In the solution, all j does is delegate two tasks to $(j-1)$ and move a single disk. The total number of transfers under j's authority is one plus twice the number under $(j-1)$'s authority. Moving a single disk takes one transfer; moving a stack of two disks therefore takes $1+2\cdot 1 = 3$ transfers, moving a stack of three disks takes $1+2\cdot 3 = 7$ transfers, moving a stack of four disks takes $1+2\cdot 7 = 15$ transfers, and so on. It is left as an exercise to guess a formula for the number of transfers required to move a stack of n disks, and to prove this guess is correct by mathematical induction.

It should be clear why recursive definitions are so useful; an immense amount of complexity can be encoded in a small set of recursive rules. Each person in the solution of the Tower of Hanoi needs to know only two trivial things, but by coordinated delegation of tasks they solve a complicated problem. However, the number of transfers needed essentially doubles with each additional disk. Suppose one disk can be moved per second. To move two disks will take at least 2 seconds, to move three disks will take at least 4 seconds, and so on (this is a lower bound, not an exact count). To move a stack of 7 disks will take more than a minute. A stack of 13 disks will take about an hour if no mistakes are made, a stack of 20 will take about a week, a stack of 35 is well beyond a single human lifetime, and a stack of 60—at one transfer per second—would take considerably longer than the universe is believed to have existed. The Brahmin priests of the legend will not

complete their task before the earth is destroyed by well-understood astronomical phenomena. □

Properties of Addition

As a final application of mathematical induction, let us see how properties of addition follow from the axioms for the natural numbers. The arguments are relatively involved; they use nothing more than induction, but the choices of inductive statements are sometimes delicate. At least a skimming is recommended, though the proof of Theorem 2.6 may be skipped on a first reading.

Theorem 2.6. *Addition is associative and commutative; that is, if m, n, and ℓ are natural numbers, then $m + (n + \ell) = (m + n) + \ell$ and $n + m = m + n$.*

Proof. Equation (2.2) says (with different letters) that

$$(*) \qquad p + (q + 1) = (p + q) + 1 \qquad \text{for all } p \text{ and } q \in \mathbf{N};$$

associativity for $\ell = 1$ is built into the definition of addition. To prove associativity in general, consider the statement

$$A(\ell): \qquad m + (n + \ell) = (m + n) + \ell \qquad \text{for all } m \text{ and } n \in \mathbf{N}.$$

The base case is given by $(*)$. Assume $A(k)$ is true for some $k > 1$; then for each choice of m and $n \in \mathbf{N}$,

$$\begin{aligned} m + \big(n + (k + 1)\big) &= m + \big((n + k) + 1\big) && (*)\text{: } p = n,\ q = k \\ &= \big(m + (n + k)\big) + 1 && (*)\text{: } p = m,\ q = n + k \\ &= \big((m + n) + k\big) + 1 && \text{by } A(k) \\ &= (m + n) + (k + 1) && (*)\text{: } p = m + n,\ q = k \end{aligned}$$

Thus, $A(k)$ implies $A(k+1)$; because $A(1)$ is true, Theorem 2.2 says $A(\ell)$ is true for all $\ell \in \mathbf{N}$, that is, addition is associative.

Commutativity is proven by a double application of induction; first show that $n + 1 = 1 + n$ for all $n \in \mathbf{N}$ (by induction on n), then prove that $n+m = m+n$ for all m, $n \in \mathbf{N}$ (by induction on m). Associativity is used several times. Consider the statement

$$P(n): \qquad n + 1 = 1 + n.$$

The base case $P(1)$ says the successor of 1 is the successor of 1 (or $1 + 1 = 1 + 1$), which is obviously true. Now assume $P(k)$ is true for some $k \in \mathbf{N}$; we wish to prove $P(k + 1)$. But

$$\begin{aligned} (k+1)+1 &= (1+k)+1 && \text{inductive hypothesis } P(k) \\ &= 1+(k+1) && \text{by associativity.} \end{aligned}$$

This proves $P(k + 1)$, so by induction $P(n)$ is true for all $n \in \mathbf{N}$. Now consider the statement

$$C(m): \qquad n+m = m+n \qquad \text{for all } n \in \mathbf{N}.$$

The *infinite sequence of assertions* $\{P(n) : n \in \mathbf{N}\}$ is exactly the base case $C(1)$. Assume $C(k)$ is true for some natural number k, namely that $n + k = k + n$ for all $n \in \mathbf{N}$. Then for all $n \in \mathbf{N}$,

$$\begin{aligned} n+(k+1) &= (n+k)+1 && \text{associativity} \\ &= (k+n)+1 && \text{inductive hypothesis, } C(k) \\ &= k+(n+1) && \text{associativity} \\ &= k+(1+n) && \text{by } P(n) \text{ above} \\ &= (k+1)+n, && \text{associativity} \end{aligned}$$

that is, $C(k + 1)$ is true. By induction, $C(m)$ is true for all m, so addition is commutative. □

A substantial amount of work has been expended simply to place grade-school arithmetic on a set-theoretic footing, though the tools developed—recursive definition and mathematical induction—are worth the effort. Observe that the symbols "2," "4," "+," and "=" have been defined in terms of sets, and that the theorem "2+2 = 4" has essentially been proven.

Multiplication and Exponentiation

Just as addition of natural numbers was defined to be iterated succession, multiplication is defined to be iterated addition: $2 \times 3 = 2 + 2 + 2$ ('add 2 to itself' 3 times), for example. Precisely, define $m \times 0 = 0$ for all $m \in \mathbf{N}$, then for $n \geq 0$ define

$$m \times (n+1) = (m \times n) + m \qquad \text{for all } m \in \mathbf{N}. \tag{2.4}$$

Note that the distributive law is built into the recursive specification of multiplication. An excellent exercise is to mimic the proof of Theorem 2.6, showing that multiplication is associative and commutative, and distributes over addition.

Going one step further, exponentiation is iterated multiplication: $2^3 = 2 \times 2 \times 2$ ('multiply 2 by itself' 3 times).[3] Precisely, we set $m^0 = 1$ for all $m > 0$, then define, for $n \in \mathbf{N}$,

$$m^{n+1} = (m^n) \times m \qquad \text{for } m > 0. \tag{2.5}$$

(We make the special definition $0^n = 0$ for all $n > 0$; the expression 0^0 is not defined.) You should recursively expand these definitions and write the algorithms as pseudocode as an exercise. Observe that exponentiation is immensely complicated when expressed in terms of successorship.

Addition and multiplication are commutative and associative, and it would not be unreasonable to suspect this is true of the operations obtained by successive iteration of them. However, exponentiation is *neither commutative nor associative*! In fact, aside from the trivial cases $m = n$, there is only one pair of natural numbers (2 and 4) for which exponentiation commutes. It is amusing to ponder why $2^4 = 4^2$, and why these numbers are exceptional in this regard.

Relations

Let X be a set. Recall that $X \times X$ is the set of ordered pairs of elements of X.

Definition 2.7 A *relation* on X is a set $R \subset X \times X$ of ordered pairs from X. If x and y are elements of X, then we say "x is related to y by R" or "xRy" if $(x, y) \in R$.

In the examples below, filled circles represent elements of R.

Example 2.8 Let $X = \mathbf{N}$, the set of natural numbers, and let $R \subset \mathbf{N} \times \mathbf{N}$ be the set of pairs (m, n) with $m < n$, see Figure 2.4. In this case, m is related to n exactly when $m < n$. □

Example 2.9 Again let $X = \mathbf{N}$, and let R be the set of pairs (m, n) for which $m + n$ is even, Figure 2.5. In this case, m is related to n when m and n have the same *parity*, that is, are both even or both odd. □

[3] If generalization comes to mind, you are ready for the *Ackerman function*!

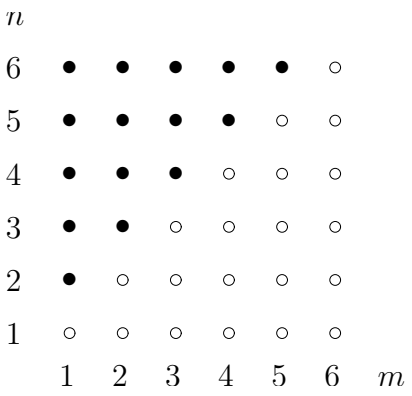


Figure 2.4: Less than.

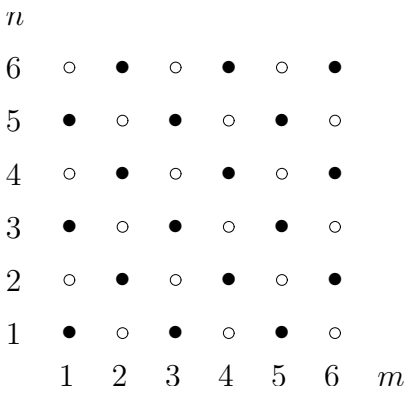


Figure 2.5: Parity.

Example 2.10 Let X be an arbitrary set. The relation of *equality* is defined by the subset $\Delta = \{(x, x) : x \in X\}$, Figure 2.6. Two elements of X are related by Δ exactly when they are equal. The subset Δ is often called the *diagonal* of $X \times X$. □

Example 2.11 Let X be a set having more than one element. The relation of *inequality* is defined by the complement of the diagonal, Figure 2.7, namely by $R = (X \times X) \setminus \Delta = \{(x, y) \in X \times X : x \neq y\}$. □

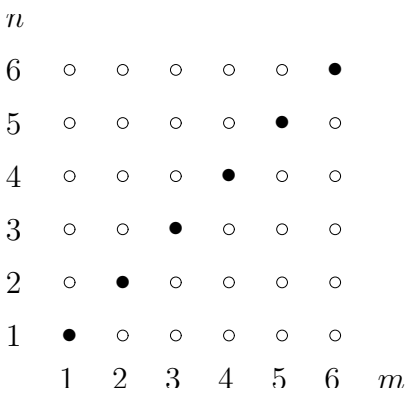


Figure 2.6: Equality.

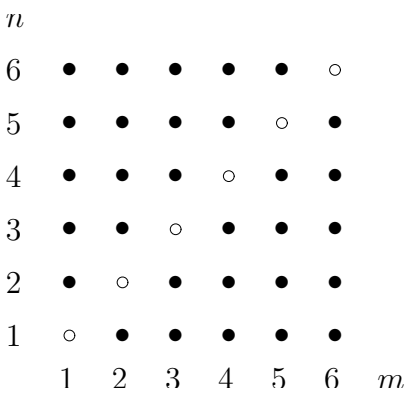


Figure 2.7: Inequality.

Definition 2.12 Let X be a set. An *equivalence relation* on X is a relation, usually denoted $\sim$, such that

- (Reflexivity) For all $x \in X$, $x \sim x$. In words, every element is related to itself.
- (Symmetry) For all x and $y \in X$, $x \sim y$ if and only if $y \sim x$. Roughly, the relation sees only whether x and y are related or not, and does not otherwise distinguish pairs of elements.
- (Transitivity) For all x, y, and $z \in X$, if $x \sim y$ and $y \sim z$, then $x \sim z$. Intuitively, the relation is all-encompassing; everything related to something related to x is itself related to x.

If $\sim$ is an equivalence relation on a set X, then there is a "partition" of X into disjoint subsets called *equivalence classes*, each consisting of elements that are mutually related by $\sim$. The equivalence class of $x \in X$ is defined by

$$[x] = \{y \in X : x \sim y\} \subset X.$$

The set of equivalence classes is denoted $X/\sim$, read X *modulo equivalence*. Intuitively, an equivalence relation "is blind to" certain distinctions; it cannot distinguish elements in a single equivalence class. The "parity" relation on $\mathbf{N}$ is an equivalence relation, with two equivalence classes: $[0] = \{\text{even numbers}\}$ and $[1] = \{\text{odd numbers}\}$. These classes can be written in (infinitely) many ways, e.g., $[1] = [3] = [329]$.

Remark 2.13 The "less than" relation is transitive, but neither reflexive nor symmetric. The relation of "inequality" is symmetric, but neither reflexive nor transitive (if $x \neq y$ and $y \neq z$, it *does not follow* that $x \neq z$ in general). The "empty" relation on a non-empty set ($R = \varnothing$; nothing is related to anything else) is symmetric and transitive (because the hypotheses are vacuous) but is not reflexive.

There is a clever (but erroneous!) argument that a symmetric and transitive relation must be reflexive: If x is related to y by a symmetric relation, then (so the argument goes) taking $z = x$ in the transitivity property shows that x is related to x. As shown by the "empty relation," the error is in assuming that every element is actually related to something. □

2.2 Integers

The natural numbers have an asymmetry; while every natural number has a successor, not every natural number has a predecessor. Consequently, equations such as $2 + x = 1$ have no solution in the set of

natural numbers. To circumvent this deficiency, we construct—using only natural numbers and operations of set theory—a larger collection of "numbers" that contains a copy of $\mathbf{N}$ and in which equations like $n_1 + x = n_2$ always have solutions (when n_1 and n_2 are numbers *of the more general type*). This larger set of numbers is the set $\mathbf{Z}$ of *integers*.[4]

In the following discussion, we speak of objects we wish to define as if they already exist. This is not logically circular, because we seek only to motivate the eventual definition, not to prove anything.

We take our cue from the equation $n_1 + x = n_2$: Given an ordered pair (n_1, n_2) of natural numbers with $n_1 \leq n_2$, there exists a unique natural number x with $n_1 + x = n_2$. Suppose we were to "define" an integer to be an *ordered pair of natural numbers*, with the idea that the pair $x = (n_1, n_2)$ corresponds to the solution of $n_1 + x = n_2$. This seems promising, because negative numbers could be realized as ordered pairs with $n_1 > n_2$; for example, the pair $(2, 1)$ would correspond to the solution of $2 + x = 1$, namely to the integer $x = -1$.

The small hitch is that many different pairs would represent the same number; $(1, 4)$, $(6, 9)$, and $(1965, 1968)$ all correspond to the number 3. In fact, two pairs (n_1, n_2) and (m_1, m_2) represent the same number exactly when $n_2 - n_1 = m_2 - m_1$, that is, when $n_2 + m_1 = n_1 + m_2$. We are therefore led to define the relation

$$(m_1, m_2) \sim (n_1, n_2) \text{ iff } n_2 + m_1 = n_1 + m_2. \tag{2.6}$$

This equation defines an equivalence relation on the set $X = \mathbf{N} \times \mathbf{N}$, as you can check, and uses nothing but the set of natural numbers and the operation of addition. We arrive at an "implementation" of the integers in terms of sets:

Definition 2.14 An *integer* is an equivalence class of $\mathbf{N} \times \mathbf{N}$ with respect to the relation (2.6).

For example, using boldface numerals to denote integers in the usual way,

$$\mathbf{2} = [(0, 2)] = [(5, 7)], \quad -\mathbf{2} = [(2, 0)] = [(7, 5)], \quad \mathbf{0} = [(0, 0)] = [(4, 4)].$$

In this construction, a *single integer* consists of infinitely many pairs of natural numbers! The equivalence class $\big[(0, n)\big] \in X$ corresponds to the natural number n, so we have succeeded in building a copy of $\mathbf{N}$ inside $\mathbf{Z}$.

[4]The abbreviation comes from German, probably from *Zahl* (for "number") or *Zyklus* (for "cycle", a reference to abstract algebra).

It remains to define addition and multiplication of integers. Think of each integer (equivalence class) as a hat containing infinitely many slips of paper, each slip having an ordered pair of natural numbers written on it. We know that if two slips (n_1, n_2) and (n_1', n_2') are drawn from a single hat, then $n_2 + n_1' = n_2' + n_1$. To add ($\oplus$) or multiply ($\odot$) two integers, pick one slip from each of the corresponding hats; say the slips are (m_1, m_2) and (n_1, n_2). The slips

$$(\dagger) \qquad \begin{aligned} (m_1, m_2) \oplus (n_1, n_2) &:= (m_1 + n_1, m_2 + n_2) \\ (m_1, m_2) \odot (n_1, n_2) &:= (m_1 n_2 + n_1 m_2, m_1 n_1 + m_2 n_2) \end{aligned}$$

each determine an equivalence class, and these will be called the *sum* and *product* of the two original slips. (For example, plug in $(4, 2)$ and $(1, 4)$; to what integers do these pairs correspond? What are their "sum" and "product" according to these formulas?) This "definition" is provisional because it is not immediate that the "sum" and "product" would have come out the same regardless of which slips were drawn from the same two hats. The claim is that while the *numbers* on the sum and product slips depend on the original slips, the *equivalence classes* determined by (†) do not change no matter how the slips are chosen. Writing the proof in detail will ensure you understand the ideas.

We should check that when a pair of integers correspond to natural numbers, the laws of arithmetic "work the same way" as the rules for adding and multiplying natural numbers. To this end, plug $(0, n)$ and $(0, m)$ into (†):

$$\begin{aligned} (0, m) \oplus (0, n) &:= (0, m + n) \\ (0, m) \odot (0, n) &:= (0, mn) \end{aligned}$$

Make sure you understand why these equations say that "arithmetic of natural numbers regarded as integers works just like arithmetic of natural numbers". Having made this observation, it is safe to use the ordinary symbols "$+$" and "$\cdot$" when adding or multiplying integers.

More illuminating is the process that leads us to discover (†). The approach is to write down the properties we want our new operations to have, then to express the operations in terms of our definition of integers. Because we are not proving anything, there is no need for rigor. The slip (m_1, m_2) represents the integer normally called $m_2 - m_1$, and similarly for (n_1, n_2). The sum and product of $m_2 - m_1$ and $n_2 - n_1$, in ordinary notation, are

$$m_2 + n_2 - (m_1 + n_1) \quad \text{and} \quad m_1 n_1 + m_2 n_2 - (m_1 n_2 + n_1 m_2).$$

Translating into the language of pairs gives (†). Observe that the rules (†) are clearly commutative, and that + is associative as an operation on integers; multiplication is also associative, but the proof requires a short calculation.

Negative numbers were resisted by medieval philosophers as meaningless. This viewpoint is expressed facetiously in a modern joke:

> A physicist, a biologist, and a mathematician see two people go into a house. Later three people come out. The physicist thinks, "The initial observation was in error." The biologist thinks, "They have reproduced." The mathematician thinks, "If exactly one person enters the house, it will be empty again."

Of course, it is not that negative numbers are meaningless (or worse, in contradiction with set theory!), but that they cannot be used indiscriminately to model real-world situations. Using common sense, we know that the house did not start off empty.

Properties of the integers

We have "implemented" integers as equivalence classes of pairs of natural numbers, but are mostly interested in their abstract arithmetic and order properties, to which we now turn. The set $\mathbf{Z}$ of integers, together with the operation of addition, forms a mathematical structure called a (*commutative*) *group*. For future use, an abstract definition is given below. Definition 2.15 formalizes the idea that the sum of two integers is an integer, and that addition obeys certain properties.

A basic concept is that of a *binary operation* on a set G. Loosely, a binary operation $\oplus$ on G is a way of "combining" two elements x and y of G to get an element $x \oplus y$. The ordinary sum of two integers is a binary operation on $\mathbf{Z}$, as is the ordinary product.

Definition 2.15 Let G be a set, and let $\oplus$ be a binary operation on G. The pair $(G, \oplus)$ is a *commutative group* if the following are satisfied:

A.1 (Associativity) $(x \oplus y) \oplus z = x \oplus (y \oplus z)$ for all x, y, and z in G.

A.2 (Neutral element) There exists an $e \in G$ such that $x \oplus e = e \oplus x = x$ for all $x \in G$.

A.3 (Existence of inverses) For each $x \in G$, there exists an element y such that $x \oplus y = e$.

A.4 (Commutativity) $x \oplus y = y \oplus x$ for all x, y in G.

Our immediate interest is the situation $G = \mathbf{Z}$, the set of integers, for which $\oplus = +$ is ordinary addition. However, the concept of a commutative group arises in other contexts, and the unconventional "$\oplus$" is meant to remind you that in general a group operation need not be literal addition.

To discuss the additive group of integers, we switch to the more conventional "+" symbol for the binary operation, and write "e" as "0". It is customary to denote the additive inverse of n by $-n$, but for now the latter must be regarded as a *single symbol*, not as the product of n and -1. The operation of *subtraction* is merely addition of an additive inverse; observe, however, that subtraction is neither associative nor commutative!

The integers are not characterized by Axioms A1–A.4; that is, there are commutative groups that are not abstractly equivalent to the group of integers under addition. Further study of these issues belongs in a course on abstract algebra.

Axioms A.1–A.4 have some simple but useful consequences. Two of them are proven here, both for illustration and to emphasize that they are not explicitly part of the definition.

Theorem 2.16. *The neutral element in* $\mathbf{Z}$ *is unique, and every element of* $\mathbf{Z}$ *has a unique additive inverse.*

Proof. Assume 0 and $0'$ are neutral elements. Because 0 is a neutral element, Axiom A.2 with $a = 0'$ implies $0' = 0 + 0'$. Reversing the roles of 0 and $0'$ implies that $0 = 0 + 0'$; thus $0 = 0'$ as claimed.

Now suppose m and m' are inverses of n, that is, $n+m = n+m' = 0$. Then

$$m = m + (n + m') = (m + n) + m' = m'$$

by Axioms A.2, A.1, and A.2 respectively. □

Axiom systems, such as A.1–A.4, play two roles. They can be assumed outright and taken as a starting point for an abstract theory (of commutative groups, in this case). Alternatively, axioms can be regarded as *theorems* that must be established in the context of some construction. The latter viewpoint is used to prove that axiom systems are consistent. (The unproven—and unprovable—assumption is that set theory itself is logically consistent.) We shall mostly adopt the former viewpoint from now on. Lurking behind the scenes are successive *constructions* of the rational, real, and complex number systems.

2.3 Rational Numbers

Intuitively, a *rational number* p/q is a quantity that "when added to itself q times" gives p. Alternatively, p/q consists of p units of "currency" $1/q$, where q units add up to unity. If k is a positive integer, then p/q and kp/kq represent the same number, because the currency is smaller by a factor of k but there are k times as many units. Logical consistency demands that $kp/kq = p/q$ even when k is negative.

Every rational number can be expressed as a quotient of integers in infinitely many ways; for example, $1/2 = 5/10 = (-3)/(-6)$. A quotient p/q is *in lowest terms* if $q > 0$, and if p and q have no common factor larger than 1. An integer is regarded as a rational number with denominator 1. Every rational number has a *unique* representation in lowest terms.

The rational numbers do for multiplication what the integers did for addition: They allow solution of arbitrary equations $qx = p$ for $q \neq 0$ with p and q integers, and as a fringe benefit such equations are automatically solvable even when p and q are rational (with $q \neq 0$). If $q = 0$, then the expression "$p/0$" ought to stand for the solution of the equation $0x = p$. However, this equation has *no* solution if $p \neq 0$, while *every* rational number is a solution when $p = 0$. For this reason, we make no attempt to define the quotient $p/0$ as a rational number. Informally, division by zero is undefined.

If two rational numbers—say $2/5$ and $5/12$—are to be added, they must be "converted" into a common currency. This is accomplished by cross multiplication, which in this case represents the given numbers as $24/60$ and $25/60$. They can now be added (or subtracted) in the obvious way. This reasoning lies behind the general formula

$$\frac{p_1}{q_1} + \frac{p_2}{q_2} = \frac{p_1 q_2 + p_2 q_1}{q_1 q_2}, \tag{2.7}$$

which would be used to motivate the analogue of equation (†) in the construction of $\mathbf{Q}$ from $\mathbf{Z}$.

If integers are viewed as lengths (multiples of some unit length on a number line), then rational numbers can also be viewed as lengths; arbitrary rational lengths can be constructed with a straightedge and compass. It is important to develop good intuition about the way the set of rational numbers is situated on a number line. To this end, fix a positive integer q and consider the set $\frac{1}{q}\mathbf{Z}$ of integer multiples of $1/q$, namely the set of rational numbers of the form p/q, Figure 2.8. The

set $\frac{1}{q}\mathbf{Z}$ is a "scaled copy" of $\mathbf{Z}$, with spacing $1/q$, and the union of all these sets is $\mathbf{Q}$, cf. Figure 2.9.

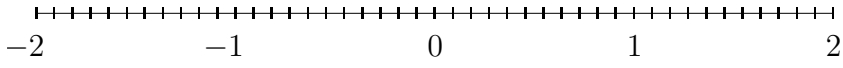


Figure 2.8: The set of rational numbers with denominator q.

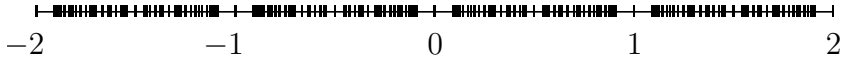


Figure 2.9: The set of rational numbers with denominator at most q.

It is not immediately clear whether or not *every* point on the number line is represented by a rational number, but Figures 2.8 and 2.9 should make it clear that every point on the line is "arbitrarily close to" a rational number. If we take the Ancient Greek idea of points on the number line as *distances*, then plausibly every distance is a ratio of integers. However, though the Pythagoreans had built their philosophy on ideas of integer ratios and harmonies, they eventually discovered—to their considerable dismay, according to legend—that the diagonal of a unit square cannot be expressed as a ratio of integers, see Theorem 1.3. If distances are to be modeled as numbers, then the set of rational numbers is inadequate.

Unfortunately, it is not clear how to describe or quantify the "gaps between rational numbers", particularly if we cannot make external reference to the number line (which, after all, we have not defined in terms of set theory). Suitable descriptions were found by R. Dedekind and G. Cantor—independently and by entirely different methods—in 1872. To give an indication of their constructions, we must first study the set $\mathbf{Q}$ of rational numbers in more detail.

Arithmetic and Order Properties of Q

The set of rational numbers together with the operation of addition, equation (2.7), forms a commutative group (see A.1–A.4 above): Addition is commutative and associative, there is a neutral element for addition, and every element has an additive inverse. The group $(\mathbf{Q}, +)$ is more complicated than the group $(\mathbf{Z}, +)$, however. Specifically, $\mathbf{Z}$ is

"generated by" the element 1, in the sense that every element of $\mathbf{Z}$ is obtained by repeatedly adding 1 (or its additive inverse) to itself. By contrast, equation (2.7) implies that there is *no finite set* of generators for $(\mathbf{Q}, +)$: If $a_1, \ldots, a_n$ are rational numbers with $a_i = p_i/q_i$ in lowest terms, then taking sums and differences cannot generate numbers whose denominator is larger than $N = q_1 \cdots q_n$, the product of the individual denominators. Consequently, the set of rational numbers generated by the a_i's has *bounded denominator*, and therefore cannot be all of $\mathbf{Q}$, see Figure 2.9.

The set of rational numbers has another arithmetic operation, multiplication, defined by

$$\frac{p_1}{q_1} \cdot \frac{p_2}{q_2} = \frac{p_1 p_2}{q_1 q_2}. \tag{2.8}$$

Let $\mathbf{Q}^\times$ denote the set of *non-zero* rational numbers. Multiplication combines two elements of $\mathbf{Q}^\times$ to give a third; the order of the factors has no bearing on the result, and the number 1 acts as a neutral element. Finally, if $r = p/q$ is a non-zero rational number, then $r^{-1} = q/p$ is also a non-zero rational, which is clearly a multiplicative inverse of r. The set $\mathbf{Q}^\times$ endowed with the operation of multiplication satisfies A.1–A.4: $(\mathbf{Q}^\times, \cdot)$ is a commutative group.

If $\mathbf{Q}$ were merely endowed with two separate group operations, then rational arithmetic would be relatively uninteresting. However, these operations are connected by the *distributive law*

$$a \cdot (b + c) = a \cdot b + a \cdot c \qquad \text{for all } a, b, \text{ and } c \in \mathbf{Q}.$$

The analogue with multiplication on the right follows by commutativity of multiplication. (Observe the asymmetry that addition does not distribute over multiplication.)

Much of elementary arithmetic can be deduced from the group axioms for addition in $\mathbf{Q}$, the group axioms for multiplication in $\mathbf{Q}^\times$, and the distributive law. A couple of examples are given here to illustrate the way the axioms are used to prove more familiar-looking properties. The second claim below, which looks tautological at first glance, asserts that the additive inverse of a is equal to the product of a and -1. (If it were not, then the notation "$-a$" would be extremely misleading!)

Theorem 2.17. *If $a \in \mathbf{Q}$, then $a \cdot 0 = 0$ and $-a = a \cdot (-1)$.*

Proof. The first assertion is proven by noticing that $0+0 = 0$ (definition of neutral element applied to $a = 0$), so multiplying both sides by a

and using the distributive law gives $a \cdot 0 + a \cdot 0 = a \cdot 0$. Adding the inverse of $a \cdot 0$ implies $a \cdot 0 = 0$.

By Theorem 2.16, $-a$ is the *unique* rational number such that $a + (-a) = 0$. If we can show that $a \cdot (-1)$ has this property, then the theorem will be proved. But we have

$$\begin{aligned} a + \big(a \cdot (-1)\big) &= \big(a \cdot 1\big) + \big(a \cdot (-1)\big) && \text{Definition of neutral element} \\ &= a \cdot \big(1 + (-1)\big) && \text{Distributive Law} \\ &= a \cdot 0 && \text{Definition of additive inverse,} \end{aligned}$$

and this is equal to 0 by the first part of the theorem. □

Abstract structures similar to $\mathbf{Q}$ arise frequently enough to warrant a special name:

Definition 2.18 A *field* $(\mathbf{F}, +, \cdot)$ consists of a non-empty set $\mathbf{F}$ and two associative operations $+$ and $\cdot$ such that

F.1 $(\mathbf{F}, +)$ is a commutative group, with neutral element $\mathbf{0}$.

F.2 The set $\mathbf{F}^\times = \mathbf{F} \setminus \{\mathbf{0}\}$ of non-zero elements of $\mathbf{F}$ is a commutative group under multiplication, with neutral element $\mathbf{1} \neq \mathbf{0}$.

F.3 Multiplication $\cdot$ distributes over addition $+$.

By customary laziness, mathematicians say "the field $\mathbf{F}$" when the operations are understood. Observe that $(\mathbf{Z}, +, \cdot)$ is *not* a field; the non-zero integer 2 has no multiplicative inverse. The study of general fields belongs to abstract algebra, and is not pursued in detail here. The axioms are given to demonstrate the conceptual economy of abstraction: A result like Theorem 2.17, which can be proven with nothing but the axioms of a field, is valid for an arbitrary field.

Finite Fields

Remarkably, there exists fields having finitely many elements. These appear in various parts of mathematics, as well as in "public-key encryption". When you view a "secure web site", whose URL begins `https://`, your web browser uses finite field arithmetic to send and receive data securely. The "number of bits" (at this writing, usually 128) is related to the number of elements of the field. The larger the field, the more

difficult the cipher is to break without the decoding key, but the slower the data transfer.

Example 2.19 A field must contain at least two elements, because we require that $\mathbf{0} \neq \mathbf{1}$. In a finite field, successive sums of $\mathbf{1}$ cannot all be distinct, so there must be a smallest positive integer p for which

$$\mathbf{1} + \cdots + \mathbf{1} = \mathbf{0} \qquad (p \text{ summands}).$$

The field axioms imply that p is a prime number; for instance, p cannot be 6, because then $\mathbf{1} + \mathbf{1}$ and $\mathbf{1} + \mathbf{1} + \mathbf{1}$ would be non-zero elements whose product is $\mathbf{0}$. Further considerations imply that the number of elements in a finite field—its *order*—must be a power of p. It is possible to construct fields of order p^k, proving their existence. Thus there exist fields of order 2, 3, $4 = 2^2$, 5, 7, $8 = 2^3$, $9 = 3^2$, 11, etc.; and there do *not* exist fields of order $6 = 2 \cdot 3$, $10 = 2 \cdot 5$, $12 = 2^2 \cdot 3$, and so forth.

Fields of prime order are easy to describe; for $p = 5$, say, take $\mathbf{F}_5 = \{0,\ 1,\ 2,\ 3,\ 4\}$ (the set of remainders upon division by 5) and define the operations to be the usual ones, except that the sum or product is divided by 5 and the remainder taken. Thus in $\mathbf{F}_5$, $2+3 = 0$ (the ordinary sum is 5, which leaves remainder 0 on division by 5) and $2 \cdot 3 = 1$ (the product is 6, which leaves remainder 1). Arithmetic in a finite field can look strange. The two examples just given can legitimately be written $-2 = 3$ (or $-3 = 2$) and $1/2 = 3$ (or $1/3 = 2$). Of course, the symbols 2 and 3 represent elements of $\mathbf{F}_5$, not rational numbers. It would be less provocative to write $-[2] = [3]$ and $[2]^{-1} = [3]$.

The construction of a field with $4 = 2^2$ elements (more generally, having p^k elements with p prime and $k > 1$) is not discussed here. It is possible for a field to have infinitely many elements, yet satisfy $\mathbf{1}+\cdots+\mathbf{1} = \mathbf{0}$ for finitely many summands. □

The Finite Geometric Series

This section presents a useful calculation that can be performed in an arbitrary field, and which furnishes a nice example of mathematical induction. If you prefer, regard the field below as $\mathbf{R}$.

Example 2.20 Let a and r be elements of a field, and let $n \in \mathbf{N}$. The expression

$$\sum_{j=0}^{n} a\, r^j = a + a\, r + a\, r^2 + a\, r^3 + \cdots + a\, r^n \tag{2.9}$$

is called a *geometric series*, and is characterized by the ratio of each pair of consecutive terms being the same (namely r). If $r = 1$, then the series consists of $n + 1$ terms, each equal to a, so the sum is $a(n + 1)$. (In a general field, this may be 0 even if $a \neq 0$!) When $r \neq 1$, the closed form of the geometric series can be guessed by multiplying the sum by $1 - r$ and "telescoping" the terms:

$$\begin{aligned}
&(1-r)(a + a\,r + a\,r^2 + a\,r^3 + \cdots + a\,r^n) \\
&= (a + a\,r + a\,r^2 + a\,r^3 + \cdots + a\,r^n) \\
&\qquad - (ar + a\,r^2 + a\,r^3 + \cdots + a\,r^n + a\,r^{n+1}) \\
&= a - ar^{n+1} = a\,(1 - r^{n+1}).
\end{aligned}$$

This argument, as with many that contain an ellipsis or the words "and so on," can be made precise with mathematical induction. (As with the sum of odd integers, the informal argument was needed to guess the answer!) Consider the statement

$$P(n): \qquad (1-r)\sum_{j=0}^{n} a\,r^j = a\,(1 - r^{n+1}).$$

When $n = 0$, this is the tautology $a\,(1 - r) = a\,(1 - r)$. Assume inductively that $P(k)$ is true for some $k \in \mathbf{N}$. The truth of $P(k + 1)$ is deduced by adding the term $a\,r^{k+1}(1 - r)$ (corresponding to $j = k + 1$) to both sides of $P(k)$ and using algebra. This is left as an exercise.

When $r = 1$, $P(n)$ merely asserts that $0 = 0$, but the informal argument given above (which could be supplemented by an induction proof) evaluates the sum. Thus

$$\sum_{j=0}^{n} a\,r^j = \begin{cases} a\,\dfrac{1 - r^{n+1}}{1 - r} & \text{if } r \neq 1, \\[2ex] a\,(n+1) & \text{if } r = 1. \end{cases} \tag{2.10}$$

Remarkably, the left-hand side of this equation is defined by a single sum, while the closed form requires two cases; the special definition at $r = 1$ occurs exactly when the expression for $r \neq 1$ becomes $0/0$. We have found a context in which the expression "$0/0$" can be ascribed meaning! □

Ordered Fields and Absolute Value

The set of rational numbers is "ordered" in an abstract sense, see Definition 2.21. The order relation on $\mathbf{Q}$ is the key to constructing the set of real numbers. Because order properties are fundamental to calculus, we introduce them in a general setting.

Definition 2.21 An *ordered field* is a field $(\mathbf{F}, +, \cdot)$ together with a subset $P \subset \mathbf{F}$ satisfying the following conditions.

O.1 (Trichotomy) For every $a \in \mathbf{F}$, *exactly one* of the following is true: $a \in P$, $-a \in P$, or $a = 0$.

O.2 (Closure under $+$) If a, $b \in P$, then $a + b \in P$.

O.3 (Closure under $\cdot$) If a, $b \in P$, then $a \cdot b \in P$.

An element of P is a *positive number*. If $-a \in P$, then a is *negative*.

Informally, O.1 says that every number is either positive, negative, or zero. O.2 and O.3 say that sums and products of positive numbers are positive. The *order relation* $<$ associated to a set P of positive numbers is defined by

$$x < y \quad \text{iff} \quad y - x \in P.$$

A field may or may not admit an order relation. A finite field never admits an order operation (why?), nor does a field in which -1 has a square root. To see the latter, observe first that in an ordered field, if $-a$ and $-b$ are negative (so that a and b are positive), then Theorem 2.17 implies that

$$(-a)(-b) = (a)(-1)(b)(-1) = (-1)^2(a \cdot b) = a \cdot b,$$

which is positive by O.3. (This simple algebraic fact caused heated dispute among medieval philosophers. Axiomization and calculation have certain advantages over intuitive arguments.) In particular, in an ordered field, the square of every non-zero element is positive. This implies that $\mathbf{1} = \mathbf{1}^2$ is positive, and $-\mathbf{1}$ is negative, in every ordered field. Thus, as claimed, if there is a square root of -1 in $\mathbf{F}$, then $\mathbf{F}$ does not admit an order relation.

Proposition 2.22. *In the rational field, there is a unique set $P \subset \mathbf{Q}$ satisfying O.1–O.3.*

Proof. Suppose $P \subset \mathbf{Q}$ satisfies O.1–O.3. As shown above, $1 \in P$. Thus every natural number $q \neq 0$ is in P by O.2. The reciprocal $1/q$ is also positive, since otherwise the equation $q \cdot (-1/q) = -1$ would contradict O.3. Finally, if p and q are non-zero natural numbers, then $p/q = p \cdot (1/q)$ is positive. We claim that

$$P = \{p/q \in \mathbf{Q} \mid p,\ q \text{ positive integers with no common factor}\} \tag{2.11}$$

We have shown that if $P \subset \mathbf{Q}$ satisfies O.1–O.3, then P is the set in (2.11), so there is at most one choice for P. But for this choice of P, each of O.1–O.3 is clear; see Exercise 2.11. □

Let x and y be elements of an ordered field. The inequality $x < y$ ("x is less than y") is also written $y > x$ ("y is greater than x"). It is convenient to write "$x \leq y$" instead of "$x < y$ or $x = y$." By trichotomy, $x \leq y$ is the negation of $y < x$.

Conventionally, a larger number lies farther to the right on the number line. Figure 2.9 makes it clear that though the set of rational numbers is ordered, rational numbers are *not* strung along the number line like beads; for each rational number x, there is no "next" rational number. In particular, *there is no smallest positive rational number.*

An order relation $<$ on a field $\mathbf{F}$ determines the set $P = \{x \in \mathbf{F} \mid 0 < x\}$. Thus an ordered field may be denoted either by $(\mathbf{F}, +, \cdot, P)$ or by $(\mathbf{F}, +, \cdot, <)$. This somewhat unwieldy notation emphasizes all the structure of an ordered field: A set of elements, two binary operations (subject to axioms), and a set of "positive" elements (subject to further axioms). In all, we have listed twelve axioms for an ordered field: Four group axioms for addition, four more for multiplication, one for the distributive law, and three order axioms.

The order axioms O.1–O.3 are low-level instructions. In practice, we want to manipulate inequalities as fluently as equalities. Theorem 2.23 successively gives rules for adding and multiplying inequalities by fixed numbers, for taking reciprocals of an inequality, and for adding and multiplying two inequalities (of *non-negative* numbers). Each of these properties is deduced from the order axioms by one of a few standard arguments. This list is not all-encompassing, but the proof illustrates the most important ideas, and other similar properties should provide easy exercises.

Theorem 2.23. *Let a, b, c, and d be elements of an ordered field.*

(i) (*Transitivity of $<$*) *If $a < b$ and $b < c$, then $a < c$.*

(ii) *If* $a < b$, *then* $a + c < b + c$ *and* $-a > -b$.

(iii) *If* $a < b$ *and* $0 < c$, *then* $ac < bc$.

(iv) *If* $c < d < 0 < a < b$, *then* $1/d < 1/c < 0 < 1/b < 1/a$.

(v) *If* $0 < a < b$ *and* $0 < c < d$, *then* $0 < a+c < b+d$ *and* $0 < ac < bd$.

The analogous assertions hold if "$<$*" is replaced by "*$\leq$*" in* (i), (ii), (iii), *and* (v).

Proof. Property (i) is a restatement of O.2; $a < b$ means $b - a \in P$, while $b < c$ means $c - b \in P$. By O.2, $c - a = (c - b) + (b - a) \in P$, so $a < c$.

(ii) Suppose $a < b$, that is, $b - a \in P$. Using the field axioms, this is re-written as $(b + c) - (a + c) \in P$, which by definition means $a+c < b+c$. Similarly, $b-a = (-a)-(-b) \in P$, which means $-b < -a$.

To prove (iii), observe that by assumption $b - a$ and c are in P. By O.3, their product $(b - a)c = bc - ac$ is in P, which means $ac < bc$.

(iv) For reciprocals, note that if $x > 0$, then $1/x > 0$ (since otherwise $-1 = x(-1/x)$ would be positive by O.3); similarly, if $y < 0$, then $1/y < 0$. Assume that $0 < a < b$. Then $0 < ab$ by O.3, so $0 < 1/(ab)$ as just shown, and applying (iii) with $c = 1/(ab)$ proves (iv).

Property (v) is a consequence of (i)–(iii): Under the hypotheses, $a + c < b + c < b + d$ and $ac < bc < bd$.

Finally, if $<$ is replaced by $\leq$, then consider separately the two cases $a < b$ and $a = b$. In the first case the proof already given implies the desired result (since, e.g., $-a > -b$ trivially implies $-a \geq -b$). If $a = b$, the conclusions are obvious. □

Let $(\mathbf{F}, +, \cdot, <)$ be an ordered field. The *absolute value* of $a \in \mathbf{F}$, denoted $|a|$, is defined by

$$|a| = \begin{cases} a \text{ if } a \geq 0 \\ -a \text{ if } a < 0 \end{cases} \tag{2.12}$$

Trivially, $0 \leq |a| = |-a|$ for all a. The quantity $|a - b| = |b - a|$ can be interpreted as the *distance* between a and b on the number line, and this accounts for its importance in analysis.

Roughly, taking the absolute value of a number "throws away the minus sign, if any." Note, however, that "$|-a| = a$" is generally false. In symbolic computations, $|a|$ can be replaced by either a or $-a$ as appropriate, so any assertion about absolute values can be established by checking sufficiently many cases. This is usually tedious, since the number of cases doubles for each additional absolute value symbol in the equation being checked.

Maxima and minima

Let $(\mathbf{F}, +, \cdot, <)$ be an ordered field, and let $a, b \in \mathbf{F}$. The *maximum* of a and b is simply the larger of the two numbers, and is denoted $\max(a, b)$. The *minimum* is defined similarly. An amusing and useful application of the absolute value is a pair of formulas for max and min:

Theorem 2.24. *Let a and b be elements of an ordered field. Then*

$$\max(a, b) = \frac{a + b + |a - b|}{2}, \qquad \min(a, b) = \frac{a + b - |a - b|}{2}.$$

Proof. For simplicity, write $M = \max(a, b)$ and $m = \min(a, b)$. There are two possibilities: $M = a$ and $m = b$, or $M = b$ and $m = a$. (If $a = b$, then both are true.) In either case, $M + m = a + b$: the sum of two numbers is their sum! Moreover, $M - m$ is non-negative and equal either to $a - b$ or to $b - a$; consequently, $M - m = |a - b|$. Solving for M and m proves the theorem. □

Some general properties of absolute value and distance are collected in Theorem 2.25. They will be used repeatedly in defining and working with limits.

Theorem 2.25. *Let a and b be elements of an ordered field. Then*

(i) $|a| \leq b$ *if and only if* $-b \leq a \leq b$.

(ii) $|ab| = |a| \cdot |b|$.

(iii) $|a + b| \leq |a| + |b|$ *and* $|a - b| \leq |a| + |b|$.

(iv) $|a - b| \geq \big||a| - |b|\big|$.

These inequalities are so important because they hold *simultaneously* for all choices of a and b, and therefore represent general properties of ordered fields.

Proof. The first property translates between a single absolute value inequality and a pair of ordinary inequalities. To prove it, note that $a \le |a|$ and $-a \le |a|$; thus $|a| \le b$ if and only if $a \le |a| \le b$ and $-a \le |a| \le b$, and the second of these is equivalent to $-b \le a$. Combining these gives (i).

Properties (ii) and (iii) are proven by checking cases. Because each claim is unchanged if both a and b are multiplied by -1, it is enough to assume one of the numbers is non-negative. Each claim is also unchanged if a and b are exchanged, so it suffices to assume $a \le b$. Finally, the two parts of (iii) are equivalent, since replacing b by $-b$ exchanges the two assertions. In summary, it is enough to verify (ii) and the second part of (iii) in the two cases $0 \le a \le b$ and $a \le 0 \le b$. In the notation of Theorem 2.24, $a = m$ and $b = M \ge 0$.

Property (ii) is easily verified in each case. The second part of (iii) is not much more difficult. If $a \le 0 \le b$, then $m = a = -|a|$ and $M = b = |b|$, so $|a-b| = M - m = |a| + |b|$. If instead $a = m \ge 0$, then $|a - b| = M - m \le M + m = |a| + |b|$. This proves (iii).

Property (iv) can be derived from (iii) by some noteworthy gymnastics:

$$|a| = |(a-b)+b| \le |a-b| + |b| \qquad \text{for all } a \text{ and } b,$$

and similarly $|b| \le |b-a| + |a| = |a-b| + |a|$. Subtracting $|b|$ from the first inequality and $|a|$ from the second gives

$$|a| - |b| \le |a-b| \quad \text{and} \quad |b| - |a| = -\big(|a| - |b|\big) \le |a-b|,$$

and this is equivalent to (iv) by (i). □

The assertions (iii) and (iv) are often called the *Triangle Inequality* and the *Reverse Triangle Inequality*, especially when written

$$|x - z| \le |x - y| + |y - z| \tag{2.13}$$

$$|x - z| \ge \big||x - y| - |y - z|\big| \tag{2.14}$$

for all x, y, z.

The first is exactly (iii) with $a = x - y$ and $b = y - z$, while the second is (iv) with $a = x - y$ and $b = z - y$. The names come from the interpretation of absolute values as distances; the length of a side of a triangle is no longer than the sum of the lengths of the other two sides, and is at least as long as (the absolute value of) the difference of their lengths.

Suprema

Which state has the lowest highest point? —Moxy Früvous

If y is a point on the number line, then there are rational numbers "arbitrarily close to" y, Figure 2.8. This idea of arbitrarily close approximation is perhaps *the* fundamental idea of calculus, and is key in eliminating the "gaps" in $\mathbf{Q}$.

Definition 2.26 Let $(\mathbf{F}, +, \cdot, <)$ be an ordered field. A subset $A \subset \mathbf{F}$ is *bounded above* if there exists an $M \in \mathbf{F}$ such that $x \leq M$ for all $x \in A$. The set A has a *maximum* if there is an $a \in A$ such that $x \leq a$ for all $x \in A$. With obvious modifications, we speak of sets that are *bounded below* or have a *minimum*. A set is *bounded* if it is bounded both above and below.

In an ordered field, a subset that has a maximum is not merely bounded above, but has a largest element; however, *a set can be bounded above without having a maximum.* This is slightly counterintuitive, perhaps because a finite set of numbers always has a largest element. The set of negative rational numbers is bounded above by 0, but there is no largest negative rational number. Clearly, if M is an upper bound of A and $M \leq M'$, then M' is also an upper bound of A; upper bounds are not unique. By contrast, a set has at most one maximum, for if a and a' are both maxima of A, then $a' \leq a$ (since a is a maximum) and $a \leq a'$ (since a' is a maximum), so $a = a'$.

The empty set is bounded (the condition is vacuous), though it has no maximum or minimum. Every non-empty *finite* set has both a maximum and a minimum, hence is bounded. The concepts of boundedness, maximum, and minimum are of greatest interest for *infinite* sets—those with infinitely many elements. Note that in common usage, "infinite" is often used in the sense of "unbounded" (e.g., "Cosmologists do not know if the universe is infinite"). It is important not to confuse these terms mathematically: An unbounded set is infinite,[5] but the set $\{x \in \mathbf{Q} \mid 0 < x < 1\}$ is infinite, yet bounded. Note that this set has neither a minimum nor a maximum.

Suppose $A \subset \mathbf{F}$ is a non-empty set that is bounded above. It is desirable, both in theory and practice, to find "the best possible upper bound" of A. "Better" means *smaller* since a smaller upper bound conveys more information; knowing a person is less than 6 feet tall gives more information than knowing they are less than 7 feet tall.

[5] Contrapositively, a finite set is bounded.

Consider the *set of upper bounds of* A. The "best" upper bound of A would be the minimum of the set of upper bounds, *if this minimum exists*. If it exists, this minimum (which is unique, by remarks made above) is called the *least upper bound* or *supremum* of A, and is denoted $\sup A$, see Figure 2.10.

If A has a maximum, then $\sup A = \max A$, but a set may have a supremum even if it has no maximum. The set $\mathbf{Q}^-$ of negative rational numbers has no maximum, but is bounded above by M for every $M \geq 0$. Consequently, 0 is an upper bound of $\mathbf{Q}^-$, and is the smallest upper bound; $\sup \mathbf{Q}^- = 0$.

Proposition 2.27. *Let $A \subset \mathbf{F}$ be a set that has a supremum, say $a = \sup A$. Then for every $\varepsilon > 0$, there is an element $x \in A$ with $a - \varepsilon < x$.*

Proof. Let $\varepsilon > 0$. Because $a = \sup A$ is the smallest upper bound of A, the number $a - \varepsilon < a$ is *not* an upper bound of A. In other words, there is an element $x \in A$ with $a - \varepsilon < x$. □

Roughly, the supremum of a set is arbitrarily close to some element of the set. This intuitive phrasing is a bit misleading, since two distinct numbers are never "arbitrarily close." The loophole is that the element x depends on the choice of ε in general. It is more accurate to say that $\sup A$ is arbitrarily close to the *set* A. If $a \in A$ (that is, A contains its supremum), then there is nothing to show; for each $\varepsilon > 0$, take $x = a$. Proposition 2.27 is most interesting for sets that do not contain their supremum.

On a number line, an upper bound of a set A is a point lying to the right of A. Moving the upper bound to the left (but staying to the right of A) "improves" the bound, and the supremum (if it exists) is the location past which the point cannot be moved without passing at least one point of A.

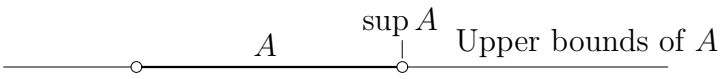


Figure 2.10: Upper bounds and the supremum.

If the set $A \subset \mathbf{F}$ is bounded above, then there are three logical possibilities:

- A has a maximum.
- A has no maximum, but has a supremum.
- A has no supremum.

Given the visual description of bounds and suprema, it may be difficult to imagine how a set that is bounded above could fail to have a supremum. School intuition emphasizes the number line, and indeed the elusive property possessed by the field of real numbers is that if A is bounded above, then the third possibility is impossible. Remarkably, a set of *rational* numbers can be bounded above but fail to have a (rational) supremum:

Proposition 2.28. *Let $A = \{x \in \mathbf{Q} \mid x^2 \leq 2\} \subset \mathbf{Q}$. The set A is non-empty and bounded above, but has no supremum.*

Proof. Since $0 \in A$, A is non-empty. Because there is no rational square root of 2, the "$\leq$" in the definition of A can be replaced by strict inequality. Let x be a positive rational number, and let $z = 2/x$. By Theorem 2.23, $x^2 > 2$ iff $z^2 < 2$, that is, $2/x \in A$ iff $x \notin A$. To show A has no supremum, we first show that the set $B = \{x \in \mathbf{Q}^+ \mid x^2 > 2\}$ has no smallest element.

Lemma 2.29. *Let $x = p/q$ be a positive rational number in lowest terms, with $x^2 > 2$, and set*

$$y = \frac{1}{2}\left(x + \frac{2}{x}\right) = \frac{1}{2}\left(\frac{p}{q} + \frac{2q}{p}\right).$$

Then y is a positive rational number with $y^2 > 2$ and $y < x$.

It is obvious that y is a positive rational number. To see that $y^2 > 2$, write $x^2 = (p^2/q^2) = 2 + \varepsilon$ (with $\varepsilon > 0$ by assumption) and calculate

$$\begin{aligned} y^2 &= \frac{1}{4}\left(\frac{p^2}{q^2} + 4 + \frac{4q^2}{p^2}\right) = \frac{1}{4}\left(2 + \varepsilon + 4 + \frac{4}{2+\varepsilon}\right) \\ &= \frac{1}{4}\left(\frac{4 + 4\varepsilon + \varepsilon^2 + 8 + 4\varepsilon + 4}{2+\varepsilon}\right) = 2 + \frac{1}{4}\cdot\frac{\varepsilon^2}{2+\varepsilon} \\ &> 2. \end{aligned}$$

Finally, $0 < \varepsilon = (p^2/q^2) - 2$, so

$$x - y = \frac{1}{2}\left(\frac{p}{q} - \frac{2q}{p}\right) = \frac{1}{2}\cdot\frac{q}{p}\left(\frac{p^2}{q^2} - 2\right) > 0,$$

proving that $y < x$ as claimed. This establishes the lemma.

Let $-B = \{x \in \mathbf{Q} \mid -x \in B\}$. Lemma 2.29 and the remarks preceding show that $\mathbf{Q}$ is the union of three *disjoint* sets: $\mathbf{Q} = -B \cup A \cup B$. It follows that a rational number x is an upper bound of A iff $x \in B$. But Lemma 2.29 implies that B has no smallest element, so A has no supremum. □

Note that A has no maximum (since the maximum would be the supremum). The proof of Lemma 2.29 gives a direct proof, too. You may find it helpful to work out the details.

2.4 Real Numbers

The concept of supremum is the correct generalization of "maximum" for infinite sets. It is not difficult to believe that much of the calculus of infinitesimals rests on the existence of suprema for all (non-empty) sets that are bounded above. For that reason, Theorem 2.30 below is central. The full proof would not contribute much of use for the rest of the book, but the basic ideas nicely illustrate several ideas discussed previously in the construction of the natural numbers and integers in terms of sets. It is best to regard Theorem 2.30 as a license to proceed; by adding a single axiom to the axioms for an ordered field, we acquire at last a number system suitable for the calculus of infinitesimals. The ordered field whose existence is asserted by Theorem 2.30 is the field of *real numbers*.

Theorem 2.30. *There exists an ordered field* $(\mathbf{R}, +, \cdot, <)$ *with the following property:*

(*Completeness*) *If* $A \subset \mathbf{R}$ *is non-empty and bounded above, then* $\sup A \in \mathbf{R}$.

This field is unique up to an isomorphism of ordered fields, and contains a copy of the rational field $(\mathbf{Q}, +, \cdot, <)$.

Proof. (Brief sketch) Dedekind's idea was to define a *single* real number to be a certain *infinite set* of rational numbers, which he called a "cut." The idea in hindsight is to associate to a real number the set of all rational numbers strictly smaller than it. To phrase this in a way that makes no reference to anything other than rational numbers, a *cut* of $\mathbf{Q}$ is a non-empty set $X \subset \mathbf{Q}$ such that

- X is bounded above and has no largest element;
- If $x \in X$ and $x' < x$, then $x' \in X$.

The beauty of this construction is multifold: If X and Y are cuts, then either $X \subseteq Y$ or $Y \subseteq X$; the order relation on $\mathbf{R}$ is induced by inclusion of sets. Furthermore, rational numbers correspond to *cuts that already have a supremum*; for example, the set $\mathbf{Q}^-$ of negative rational numbers is a cut, and corresponds to the number 0.

Completeness is easy to see, since if $A \subset \mathbf{R}$ is bounded above (namely, is a *collection of cuts* for which there is a single upper bound), then the union of these cuts is itself a cut, and is readily seen to be the supremum of A. Finally, addition is easy to define; the sum of two cuts is the set of sums obtained by adding the elements of X to the elements of Y. The one annoyance is that multiplication is slightly messy to define; taking the set of pairwise products in analogy to the definition of addition does not work, because cuts contain negative numbers of large absolute value. Once multiplication has been defined, there are many details to check, namely that addition and multiplication of cuts satisfy the field axioms, and that the order axioms are satisfied.

Cantor's construction of the reals—outlined briefly in Chapter 4—is completely different; his definition is more complicated than Dedekind's, but the field and order axioms are easier to establish. □

At risk of belaboring a point, let us take stock of what Theorem 2.30 provides. There exists a field $(\mathbf{R}, +, \cdot)$ that extends the rational number system. It is possible to compare two real numbers in the sense of the order axioms O.1–O.3. Completeness says that any putative quantity that is approximated arbitrarily closely from below by real numbers is itself a real number. It is here where the rational numbers fail, for as in Proposition 2.28 the diagonal of a unit square can be approximated arbitrarily closely by rational numbers, but is not itself rational. This deficiency is serious because the basic operation of analysis—taking a "limit"—is often accomplished by approximation from below.

The uniqueness assertion in Theorem 2.30 means that any two implementations of the axioms for a complete, ordered field are abstractly equivalent. The same cannot be said of the axioms for an ordered field: $\mathbf{Q}$ and $\mathbf{R}$ are ordered fields, but are not abstractly equivalent. For example, every positive real number has a real square root, but not every positive rational number has a rational square root.

Infima

Everything that has been said about upper bounds has a version, with suitable modifications, for lower bounds. Suppose A is a non-empty set (in some ordered field) that is bounded below. If the set of lower bounds has a maximum, then this maximum is (naturally) called the *greatest lower bound* or *infimum* of A. An easy way to see the relationship between upper and lower bounds is to consider, for a non-empty set A in a field, the set $-A = \{-x : x \in A\}$. By Theorem 2.23, multiplication by -1 reverses the order relation in an ordered field, so the negative of an upper bound of A is a lower bound of $-A$. Propositions 2.27 and 2.28 have obvious restatements for infima, and completeness can be formulated in terms of infima. Proposition 2.31 below lists a couple of elementary relations between suprema and infima. The statement should not be surprising, and the proof is left as an exercise.

Density of the rationals

Proposition 2.31. *Let $A \subset \mathbf{R}$ be non-empty. Then $\sup(-A) = -\inf A$ and $\inf A \leq \sup A$, with equality if and only if A consists of a single element.*

It is fairly clear from the construction of $\mathbf{Q}$ that $\mathbf{N}$ is not bounded above in $\mathbf{Q}$. In fact, $\mathbf{N}$ is not bounded above in $\mathbf{R}$; this fact, a special case of the *Archimedean property* of $\mathbf{R}$, is a consequence of completeness, and is of central importance in the study of "limits."

Theorem 2.32. *For every $a > 0$ and every $R \in \mathbf{R}$, there exists an $n \in \mathbf{N}$ such that $an > R$.*

Said another way, if $a > 0$, then the set $a\mathbf{N} = \{an \mid n \in \mathbf{N}\}$ is not bounded above in $\mathbf{R}$. In a vaguely Taoist metaphor, "A journey of a thousand miles (R) is taken step by step (one a at a time)."

Proof. Fix a real number $a > 0$ and suppose there were an upper bound of $a\mathbf{N}$. By completeness, there would exist a least upper bound, say R. But then $R - a$ would not be an upper bound of $a\mathbf{N}$, so there would be an $n \in \mathbf{N}$ with $an > R - a$. This in turn would imply $a(n+1) > R$, and since $n + 1 \in \mathbf{N}$ this implies R is not an upper bound of $a\mathbf{N}$, contradicting $R = \sup a\mathbf{N}$. □

The reason for stating such an "obvious" fact is that there exist ordered fields that *contain* the real field; in these fields, there exist elements that are genuinely infinite or infinitesimal, and in such a field the set $\mathbf{N}$ *is* bounded above. Such a "non-Archimedean" field is not complete; adding more elements to $\mathbf{R}$ introduces new gaps.

The Archimedean property of $\mathbf{R}$ can be used to prove a fundamental approximation property of rational and real numbers: Between two distinct real numbers, there exists a rational number. In particular, every real number is arbitrarily close to the set of rationals; we say the set of rationals is *dense* in $\mathbf{R}$. The geometric idea is simple enough: If $x < y$, then $y - x > 0$. Choose a positive integer q such that $\frac{1}{q} < y - x$. Consecutive elements of $\frac{1}{q}\mathbf{Z}$ are spaced more closely than x and y, so at least one element must lie between x and y. For future use, a formal statement and proof are given here.

Theorem 2.33. *Let x and y be real numbers with $x < y$. There exists a rational number $r = \frac{p}{q}$ such that $x < r < y$.*

Proof. First consider the case $0 < x < y$. By hypothesis, $0 < y - x$, so $0 < z := 1/(y-x)$ by Theorem 2.23. The Archimedean property implies there exists a positive integer $q > z$; thus $1/q < y - x$. Now consider the set $A = \{n \in \mathbf{N} \mid x < n/q\}$. By the Archimedean property with $R = x$ and $a = 1/q$, $A \neq \varnothing$. Also, $0 \notin A$ because $0 < x$. Property **N**.3 implies the set A has a smallest element $p > 0$. Because $p - 1 \in \mathbf{N}$ is not in A, it follows that $(p-1)/q \leq x < p/q$. But $1/q < y - x$, so

$$\frac{p}{q} = \frac{p-1}{q} + \frac{1}{q} \leq x + \frac{1}{q} < x + (y - x) = y,$$

proving that $p/q < y$.

If $x < y < 0$, we apply the argument above to the numbers $0 < -y < -x$, deducing existence of a rational r with $-y < r < x$. Theorem 2.23 says $x < -r < y$, which proves the theorem in this case. In the remaining case, $x < 0 < y$, the conclusion of the theorem is obvious. □

Corollary 2.34. *Let $x \in \mathbf{R}$, and let $\varepsilon > 0$ be arbitrary. There exists a rational number $r = \frac{p}{q}$ such that $|x - r| < \varepsilon$.*

This is an immediate consequence of Theorem 2.33: Let $y = x + \varepsilon$, for example.

Completeness and Geometry

There is a subtle point about geometry that was not fully appreciated until the 20th Century: Euclid's *Elements* was not entirely based on his axioms, but relied tacitly on completeness. Strictly speaking, many of his theorems are incorrect as stated, though of course Euclid's proofs are correct once the foundations of geometry are properly developed.

The set of real numbers may be regarded axiomatically or geometrically, but while the geometric picture is often more intuitively compelling, it will always be necessary for us to translate assertions into the language of the axioms in order to verify them. In this book, the role of geometry is to foster understanding and discovery, while the field, order, and completeness axioms support logical, deductive proof.

Axioms for the Real Field

From now on we view the field of real numbers as specified by axioms of arithmetic, order, and completeness; these contain all the information about $\mathbf{R}$ needed to develop the calculus of differentials and integrals, and provide a clean logical foundation for the rest of the book. For the record, these axioms are collected here.

Definition 2.35 The *field of real numbers* is a set $\mathbf{R}$ together with binary operations "+" and "$\cdot$" and a subset P satisfying the following conditions:

- Addition axioms: $(\mathbf{R}, +)$ is a commutative group
  - $+$ is associative: $(x+y)+z = x+(y+z)$ for all x, y, z in $\mathbf{R}$
  - There is an element 0 in $\mathbf{R}$ with $x + 0 = x$ for all x in $\mathbf{R}$
  - For every x in $\mathbf{R}$, there exists a y in $\mathbf{R}$ such that $x + y = 0$
  - The operation $+$ is commutative: $x + y = y + x$ for all x, y in $\mathbf{R}$
- Multiplication axioms: $(\mathbf{R}^\times, \cdot)$ is a commutative group ($\mathbf{R}^\times := \mathbf{R} \setminus \{0\}$)
  - The operation $\cdot$ is associative: $(xy)z = x(yz)$ for all x, y, z in $\mathbf{R}$
  - There is an element 1 in $\mathbf{R}^\times$ with $x \cdot 1 = x$ for all x in $\mathbf{R}$

  - For every x in $\mathbf{R}^\times$, there exists a y in $\mathbf{R}^\times$ such that $xy = 1$
  - The operation $\cdot$ is commutative: $xy = yx$ for all x, y in $\mathbf{R}$

- The distributive law: $x(y + z) = xy + xz$ for all x, y, z in $\mathbf{R}$
- Order axioms:
  - Trichotomy: For each x in $\mathbf{R}$, *exactly one* of $x \in P$, $-x \in P$, or $x = 0$ holds.
  - Closure under $+$: If $x \in P$ and $y \in P$, then $x + y \in P$
  - Closure under $\cdot$: If $x \in P$ and $y \in P$, then $xy \in P$
- Completeness: If $A \subset \mathbf{R}$ is non-empty and bounded above, then $\sup A \in \mathbf{R}$.

Representations of Real Numbers

In "real-world" applications (as well as in many mathematical applications) one needs some concrete way to write real numbers, similar to the way of writing rational numbers as quotients of integers. The most familiar scheme is undoubtedly decimal notation, which compactly encodes a sequence of rational numbers that furnish successively better approximations to a real number. For example, the expression $1.414213\ldots$ stands for the sequence

$$\frac{1}{1}, \frac{14}{10}, \frac{141}{100}, \frac{1,414}{1,000}, \frac{14,142}{10,000}, \frac{141,421}{100,000}, \frac{1,414,213}{1,000,000}, \ldots.$$

This notation has the minor drawback that certain pairs of expressions correspond to the same real number: $.9\overline{9} = 1$, for example.

Decimal notation is predicated on dividing by powers of 10, and arose because humans have ten fingers (themselves called *digits* in biology). For each positive integer b there is an analogous "base b" notation.[6] Arguably, the only "natural" notation is *binary*, or base 2, though it takes a longer expression to get the same amount of accuracy (the binary expression 0.000001 represents the rational number 1/64, which is larger than the decimal 0.01). These issues are explored in detail in Exercises 2.17, 2.18, and 2.19.

[6]Devotees of *The Simpsons* may recall that Homer counts in octal—base 8.

A more natural form of representation is by *continued fractions*. Every real number x can be written uniquely as

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots}}},$$

with a_0 an integer and a_k positive integers for $k \geq 1$. The algorithm for generating the integers a_k, and a few basic properties of continued fractions, are given in Exercise 2.22.

On Mathematical Constants

With the advent of electronic calculators has come a belief that certain constants are *defined* by their decimal expansion. This belief is nonsense, and should be dispelled immediately. A non-repeating decimal expansion contains an infinite amount of information unless some rule is known to find successive digits. The symbol "$\sqrt{2}$" is *not* defined to be $1.4142135623\ldots$; a definition must specify the thing being defined, but the ellipsis omits an infinite number of digits. Instead, mathematicians take a practical point of view: "$\sqrt{2}$" is a positive number whose square is 2. (This immediately raises more questions: Does such a number exist? Could more than one number have this property? How can a decimal representation be found?) Real numbers like π, e, or $2^{\sqrt{2}}$ are defined by properties, not as decimal representations; finite decimal expressions are merely rational approximations. Definitions of specific real numbers therefore usually hide surprisingly subtle theorems. That there exists a unique positive real number t with $t^2 = 2$ is a special case of Example 4.47 (an even more general result is given by Theorem 5.9), and it says something about this "obvious fact" that the proof does not occur until Chapter 4. (A proof could be given now, but the intervening material introduces concepts that will make the proof both easier to understand and more generally applicable.) There are other real numbers that we will encounter, at least some of which are surely familiar: π (the area of a unit disk, or one-half the period of the elementary trigonometric functions), e (the base of the natural logarithm), γ (Euler's constant), τ (the Golden Ratio), and so forth. Each of them is characterized by a property, and in each case there is a theorem that the property does specify a unique real number. Sometimes

one can prove that seemingly unrelated properties specify the same real number, see Chapters 12, 13, and 15.

Intervals

Among the most important sets of numbers in calculus are "intervals". Intervals exists in an arbitrary ordered field, but there is a particularly simple characterization of intervals of real numbers that does not hold for most ordered fields (such as $\mathbf{Q}$).

Definition 2.36 Let $\mathbf{F}$ be an ordered field. A set $I \subset \mathbf{F}$ is called an *interval* if, for all x and $y \in I$ with $x < y$, we have

$$x < z < y \implies z \in \mathbf{F}.$$

An interval I is said to be *bounded* (in $\mathbf{F}$) if there exists an element $R \in \mathbf{F}$ such that $-R < x < R$ for all $x \in I$, and is *open* if it contains neither a minimum nor a maximum.

In words, a set I in an ordered field is an interval if every point between a pair of points of I is also a point of I. The simplest examples of intervals are the following intervals determined by a pair of elements a and $b \in \mathbf{F}$:

$$(a, b) = \{x \in \mathbf{F} \mid a < x < b\}$$
$$[a, b] = \{x \in \mathbf{F} \mid a \leq x \leq b\}$$

You may even have seen open and closed intervals *defined* as sets of the form (a, b) or $[a, b]$, and may wonder why Definition 2.36 is so complicated. The reason is that in a general field—in fact, in every ordered field *except* the real field—there exist open intervals that are not of the form (a, b)! For example, the set

$$A = \{x \in \mathbf{Q} \mid x^2 \leq 2\}$$

is an open interval in $\mathbf{Q}$, but as we saw in Proposition 2.28, A is not of the form (a, b), regardless of the choice of a and $b \in \mathbf{Q}$. The real field is special in this regard because of the completeness property:

Proposition 2.37. *If $I \subset \mathbf{R}$ is a bounded, non-empty open interval, then there exist real numbers a and b such that $I = (a, b)$.*

Proof. Because $I \subset \mathbf{R}$ is bounded and non-empty, $a := \inf I$ and $b := \sup I$ exist. Because I is open, a and b are not elements of I, so $a < b$. We seek to establish two inclusions:

$$I \subset (a, b) \quad \text{and} \quad (a, b) \subset I.$$

The first assertion is clear; if $z \in I$, then by definition of inf and sup we have $a \leq z \leq b$. Since a and b are not elements of I, we must have $a < z < b$, so $z \in (a, b)$.

The second assertion is equally easy. Suppose $a < z < b$; we must show that there exist x and $y \in I$ with $x < z < y$. Now, $\varepsilon := b - z > 0$, so by Proposition 2.27 there exists a point $y \in I$ with $y > b - \varepsilon = b - (b - z) = z$. A similar argument shows there is an $x \in I$ with $x < z$. □

The proof makes it clear why the real field is the only ordered field in which open intervals are so easily characterized: In every other ordered field, there exists a bounded, non-empty set that has no supremum.

Square braces are used to denote inclusion of endpoints, as for the *closed* interval $[a, b] = \{x \in \mathbf{R} \mid a \leq x \leq b\}$. The *half-open* intervals $[a, b)$ and $(a, b]$ are defined in the obvious way. When depicting intervals geometrically, the convention is to use a filled dot "•" to denote an included endpoint, and an open dot "∘" or no dot to denote an excluded endpoint.

A-Notation

In the sciences, experimental results are never exact, and data are always presented with error intervals. For example, we can never say that two marks on a metal bar are exactly one meter apart in a mathematical sense, only that they are (say) between 0.9999 and 1.0001 meters apart. Indeed, real marks on a real metal bar themselves have width, so the idea that a single number could accurately represent a physical notion of distance is naive. Now, you may have come to believe that mathematics is completely precise regarding numerical issues, but this belief is also not realistic. What is the exact numerical value of π? Does your calculator tell you? Could a physical device *ever* give you all the decimals of π? In fact, how is π even *defined*, let alone calculated? These questions deserve answers (and will receive them in due course) but for the moment we wish to investigate the *inexact* aspects of numerical mathematics.

Before calculators were common, most students of science and engineering knew that $\pi \simeq 3.1416$. However, a philosophically careful student usually had a difficult time trying to glean the precise meaning of the symbol "$\simeq$", which (as everyone knows!) stands for *is approximately equal to*.

The mathematician's interpretation of the expression "$\pi \simeq 3.1416$" would do Goethe proud: $\pi = 3.1416 \pm 0.00005$. More formally,

$$3.14155 \leq \pi \leq 3.14165, \text{ or } |\pi - 3.1416| \leq 0.00005.$$

Roughly, without giving more decimal places, we cannot say more exactly what π is equal to. This way of quantifying our ignorance is so useful that we introduce special notation for it: $\pi - 3.1416 = A(0.00005)$. The expression on the right is read "a quantity of absolute value at most 0.00005".

This "A notation" allows us to express relationships of relative size succinctly; we can write $10^{A(2)} = A(100)$ when we mean, "If $|x| \leq 2$, then $|10^x| \leq 100$." Further convenience of A notation arises when we perform calculations in which several terms are uncertain:

$$\big(2.5 + A(0.1)\big) + \big(1 + A(0.03)\big) = 3.5 + A(0.13)$$

and

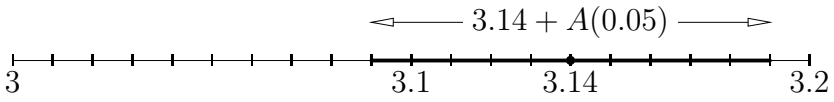
$$\begin{aligned}\big(2.5 + A(0.1)\big)\big(1 + A(0.03)\big) &= 2.5 + A(0.1) + A(0.075) + A(0.003) \\ &= 2.5 + A(0.178) = 2.5 + A(0.2).\end{aligned}$$

A notation furnishes a sort of "calculus of sloppiness".

Naturally, you must be careful not to treat A notation exactly as you would an ordinary equation; $1 = A(2)$ and $1.5 = A(2)$, but it is not true that $1 = 1.5$, nor that $A(2) = 1$. In English, this is reasonable: 1 is a number of absolute value at most 2, but a number of absolute value at most 2 is not (necessarily) 1. Similarly, $A(0.178) = A(0.2)$, but we may not conclude that $0.178 = 0.2$ or $A(0.2) = A(1.78)$. However, we *do* have $A(0) = 0$.

There is a geometric interpretation of A notation. The expression $A(0.05)$ stands for an arbitrary number in the interval $[-0.05, 0.05]$, and the expression $3.14 + A(0.05)$ stands for an arbitrary number in the interval

$$[3.14 - 0.05, 3.14 + 0.05] = [3.09, 3.19].$$



When we write $A(0.1) = A(0.2)$, it means $[-0.1, 0.1] \subset [-0.2, 0.2]$. "$A$" notation works with "variable" expressions, too:

$$x = A(x), \quad 2x = A(1 + x^2), \qquad \text{and} \qquad \frac{x^2 - 1}{x^2 + 1} = A(1).$$

The first of these is obvious, while the second is true because

$$0 \le (1 \pm x)^2 = (1 + x^2) \pm 2x \qquad \text{for all real } x,$$

so $|2x| \le 1 + x^2$. The last is left to you, see Exercise 2.7.

The Extended Real Number System

The *extended real number system* is obtained by appending two elements to $\mathbf{R}$, denoted $+\infty$ and $-\infty$. These points are not real numbers, and do not lie on the number line. By declaration,

$$-\infty < x < +\infty \qquad \text{for all } x \in \mathbf{R}.$$

If a set $A \subset \mathbf{R}$ is unbounded above, then we write $\sup A = +\infty$. Similarly, we write $\inf A = -\infty$ when A is not bounded below. If $A = \varnothing$ we agree that $\sup A = -\infty$ and $\inf A = +\infty$. (You should check that these seemingly peculiar definitions are consistent with the logic of vacuous hypotheses!) Thus every set of real numbers has a supremum and infimum in the extended reals. Note carefully that arithmetic operations with $\pm\infty$ are undefined; some expressions, such as $+\infty + (-\infty)$ *cannot* be defined in a manner that is consistent with the field axioms.

If a and b are extended real numbers, then the corresponding open interval is the set

$$(a, b) = \{x \in \mathbf{R} \mid a < x < b\}.$$

In particular, $\mathbf{R} = (-\infty, +\infty)$.

Neighborhoods

The *midpoint* of a bounded interval (a, b) (with one or both endpoints possibly included) is $(a + b)/2$, and the *radius* is $|b - a|/2$, namely one-half the length. (Compare Theorem 2.25.) It is often useful to specify

an open interval by its midpoint and radius; the *δ-interval about a* is the interval

$$N_\delta(a) := (a-\delta, a+\delta) = \{x \in \mathbf{R} : |x-a| < \delta\}$$

consisting of points whose distance to a is less than δ. The *deleted* δ-interval does not contain a:

$$N_\delta^\times(a) := N_\delta(a) \setminus \{a\} = \{x \in \mathbf{R} : 0 < |x-a| < \delta\},$$

see Figure 2.11. It is sometimes useful to regard a deleted δ-interval as a pair of intervals, $(a-\delta, a) \cup (a, a+\delta)$. Note that

$$x \in N_\delta(a) \quad \text{iff} \quad x = a + A(\delta') \text{ for some } \delta' < \delta.$$

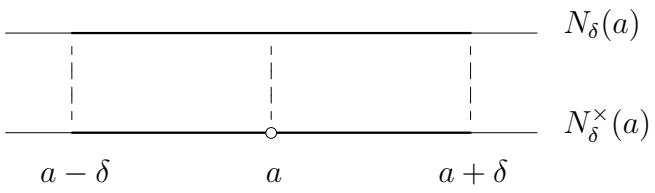


Figure 2.11: The open and deleted δ-intervals about a.

A *neighborhood* of a is a subset of $\mathbf{R}$ that contains some δ-interval about a. An open interval is a neighborhood of *each* of its points; a closed interval is not a neighborhood of its endpoints.

The set of all δ-intervals about a is used to study the behavior of "functions" near a. Remarkably, though the intersection of all the δ-intervals is $\{a\}$, the set of all such intervals captures information that cannot otherwise be seen; the set of all δ-intervals about a is a sort of "infinitesimal neighborhood" of a. The reason for considering deleted intervals is to ignore the point a explicitly, concentrating on the "infinitely near" points. (In the real field $\mathbf{R}$ this language is contradictory, but remember the Goethe quote!) A set $A \subset \mathbf{R}$ that contains some deleted δ-interval about a is said to contain all points *sufficiently close* to a.

Non-Standard Analysis

Though the traditional setting for calculus is the real field because of the historical precedent set by the Ancient Greeks, calculus can be

founded on a larger number system—the *non-standard reals*, discovered by A. Robinson—that contains "infinitesimal" numbers, namely positive numbers that are smaller than every positive real number, and "infinities" that are larger than every real number. (These infinities are considerably more subtle than the crude symbol $+\infty$ introduced for order purposes above. The field of non-standard reals is ordered and "contains a copy of $\mathbf{R}$," but is not complete; indeed, it does not satisfy the Archimedean property.) For complicated reasons, non-standard analysis occupies a position out of the mainstream of modern mathematics. Aside from the historical bias against it, there are two technical reasons for prejudice: First, it is necessary to enlarge *set theory itself* in order to construct the non-standard reals, and second, there is a theorem (the *Transfer Principle*) to the effect that every theorem about the real number system that has a "non-standard" proof also has a "real" proof. In short, a lot of expense is required, and there is no *logical* payoff; there are no new theorems that cannot be proven with "standard" techniques. Consequently, non-standard analysis is most widely known among logicians and set theorists. However, the process by which new mathematics is *discovered* is more trial-and-error than a rigid chain of logical deduction. It has been argued by I. Stewart that non-standard analysis is very useful as a conceptual tool for discovering theorems about the real numbers that would otherwise not have been found! It is not unlikely that non-standard analysis will have substantial impact on the study of physical phenomena such as the onset of turbulence in fluid flow or spontaneous symmetry breaking and phase changes.

2.5 Complex Numbers

Among the deficiencies of the real number system is the lack of general solutions to polynomial equations with real coefficients; $x^2 + 1 = 0$ has no real solution, for example. The naive attempt to remedy this situation is to assume the existence of a square root of -1 and see what logical conclusions follow. For historical reasons, a square root of -1 is denoted i, for "imaginary unit." As mentioned, the Greeks regarded "numbers" as "lengths," and there is indeed no length whose square is -1. From this point of view, i is indeed imaginary! However, the algebraic point of view is that "numbers" are merely elements of a field, and nothing prevents existence of square roots of -1 in a general field. (As we saw, an *ordered* field cannot contain such an element.)

With a bit of imagination, one is led to consider expressions $\alpha = a + bi$ with a and b real, and with arithmetic operations dictated by $i^2 = -1$ and the wish for the field axioms to hold. Complex numbers were used formally this way for over 300 years until C. F. Gauss, in the early 1800s, *defined* a complex number to be an ordered pair of real numbers, with addition and multiplication rules

$$\begin{aligned} (a,b) + (c,d) &= (a+c, b+d), \\ (a,b) \cdot (c,d) &= (ac - bd, ad + bc). \end{aligned} \tag{2.15}$$

These operations satisfy the field axioms, as may be checked by direct calculation. The reciprocal of a non-zero complex number $\alpha = a + bi$ is

$$\frac{1}{\alpha} = \frac{a - bi}{a^2 + b^2}.$$

The real number a corresponds to the complex number $(a, 0)$, and the operations (2.15) behave as expected:

$$(a,0) + (c,0) = (a+c, 0), \qquad (a,0) \cdot (c,0) = (ac, 0).$$

In words, there is a copy of $\mathbf{R}$ sitting inside $\mathbf{C}$. The pair $(0,1)$ is immediately seen to satisfy $(0,1)^2 = (-1,0)$, so the complex number system contains a square root of -1. In our modern point of view, Gauss *constructed* the field $\mathbf{C}$ of complex numbers from the field of real numbers, which proved the logical consistency of existence of $\sqrt{-1}$.

Complex numbers are represented geometrically as a number *plane*. Addition and multiplication of complex numbers have beautiful geometric descriptions, see Figure 2.12. We will prove that this description is correct in Chapter 15. Addition of a complex number α is given by the *parallelogram law*, which translates the origin to α. Multiplication by $\alpha \in \mathbf{C}$ is the operation of rotating and scaling the plane about the origin in such a way that 1 is carried to α. In particular, multiplication by -1 corresponds to reflection in the origin (i.e., one-half of a full rotation), while multiplication by $i = (0,1)$ is a counter-clockwise one-quarter rotation of the plane. This picture imbues complex numbers with an existence as "real" as that of real numbers.

There are actually *two* imaginary units, i and $-i$; by custom, i is represented by the point $(0,1)$. The *complex conjugate* of $\alpha = a + bi$ is the complex number $\bar{\alpha} = a - bi$. Geometrically, conjugate numbers are reflected across the horizontal axis. A short calculation shows that

$$\overline{(\alpha + \beta)} = \bar{\alpha} + \bar{\beta}, \qquad \overline{(\alpha \cdot \beta)} = \bar{\alpha} \cdot \bar{\beta}. \tag{2.16}$$

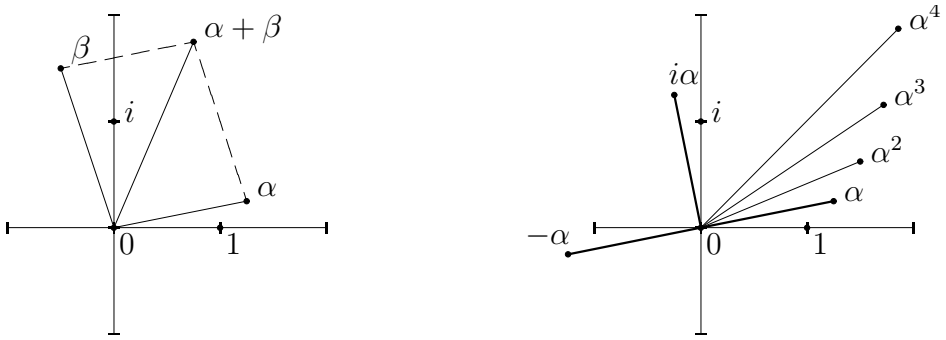


Figure 2.12: The geometry of complex addition and multiplication.

In words, the field operations work the same way after conjugating: There is no algebraic reason to identify the pair $(0, 1)$ with i rather than $-i$.

The *norm* (or *absolute value*) of a complex number is the distance to the origin, or $|\alpha| = \sqrt{\alpha\bar{\alpha}} = \sqrt{a^2 + b^2}$, and the distance between α and β is $|\alpha - \beta|$. This definition agrees with the Pythagorean theorem, and satisfies the triangle and reverse triangle inequalities:

Theorem 2.38. *For all complex numbers α and β,*

$$\big||\alpha| - |\beta|\big| \leq |\alpha - \beta| \leq |\alpha| + |\beta|.$$

The proof—which is deferred to Chapter 15, see Theorem 15.1—is not difficult if organized carefully, but is not an obvious generalization of the proof for real numbers. However, the geometric interpretation is the same, with the added bonus that "triangles" in the complex plane really are triangles.

Exercises

Exercise 2.1 The field $\mathbf{F}_2$ can be viewed as the set $\{[0], [1]\}$, where $[0]$ is the set of even integers and $[1]$ is the set of odd integers. Find the addition and multiplication tables for this field. Is there a square root of $-[1]$ in $\mathbf{F}_2$? (Hint: What *is* $-[1]$?) Using the correspondence $[0] \leftrightarrow$ False, $[1] \leftrightarrow$ True, find Boolean operations that correspond to addition and multiplication. ◇

Exercise 2.2 Recall that equation (2.2), the definition of addition, is the base case for associativity. Similarly, the definition of multiplication, equation (2.4), is the inductive base case of the distributive law. Which well-known identity has equation (2.5), the definition of exponentiation, as base case? ◇

Exercise 2.3 The discovery that two mathematical structures are "abstractly the same" can lead to new discoveries, because things difficult to see from one point of view may be easy to see from another.

Here is a whimsical example, due to Martin Gardner. Nine cards, labelled 1–9, are placed in order on a table:

1	2	3	4	5	6	7	8	9

Two players alternate taking cards. The object is to draw *three* cards that sum to 15. For example, if the first player draws 3, 8, 5, and 4 (in that order), then the cards $\{3, 4, 8\}$ constitute a win (assuming the second player did not win in the meantime). If neither player succeeds the game is a draw.

Use the 3×3 magic square

2	9	4
7	5	3
6	1	8

to show that this game is equivalent to tic-tac-toe, in the sense that there is a correspondence between winning strategies in the two games. ◇

Exercise 2.4 Let (a_i) and (b_j) be sequences, and let c be a number. Use mathematical induction to prove the following statements.

(a) $\sum_{k=0}^{n}(a_k + b_k) = \sum_{i=0}^{n} a_i + \sum_{j=0}^{n} b_j$ for all $n \geq 0$.

(b) $\sum_{i=0}^{n}(ca_i) = c\sum_{i=0}^{n} a_i$ for all $n \geq 0$.

(c) $\left(\sum_{i=0}^{n} a_i\right)\left(\sum_{j=0}^{m} b_j\right) = \sum_{i=0}^{n}\left(\sum_{j=0}^{m} a_i b_j\right)$ for all m and $n \geq 0$.

◇

Exercise 2.5 (a) Use induction to show that $1 + n \leq 2^n$ for every $n \in \mathbf{N}$, with equality iff $n = 1$.

(b) Show more generally that $1 + nr \leq (1 + r)^n$ for all $r > 0$ and all $n \geq 1$, with equality iff $n = 1$. This is a trivial consequence of the Binomial Theorem, below. Here you should do induction.

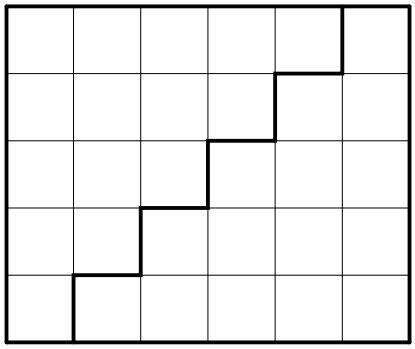
(c) Suppose $0 < r < 1$, and let $\varepsilon > 0$ be given. Prove that there is an $N \in \mathbf{N}$ such that $r^N < \varepsilon$. In words, "powers of r can be made arbitrarily small by taking the exponent sufficiently large."
Hint: If $0 < r < 1$, then there exists $x > 0$ such that $r = 1/(1 + x)$.
◇

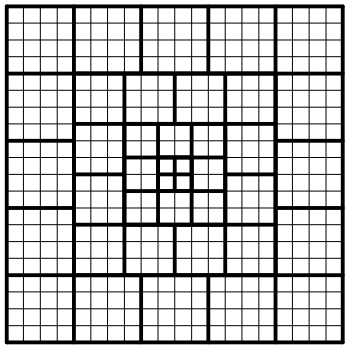
Exercise 2.6 Use induction on n to establish the following "power sum" identities:

(a) $\displaystyle\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$ (b) $\displaystyle\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1}{6}$

(c) $\displaystyle\sum_{k=1}^{n} k^3 = \frac{n^2(n+1)^2}{4} = \left(\frac{n(n+1)}{2}\right)^2$

The relationships in parts (a) and (c) may be viewed as consequences of the following diagrams:





In the second, the four squares in the center are 1; successive layers are larger, the kth layer consisting of $4k$ squares that are each $k \times k$.
◇

Exercise 2.7 Let x be a real number. Establish the following assertions.

(a) If $x = A(0.5)$, then $1 + x \geq 0.5$ and $1/(1 + x) = A(2)$.

(b) $\dfrac{x^2-1}{x^2+1} = A(1)$.

(c) If $x = A(1)$, then $x^2 = A(x)$ and $\big(1 + A(x)\big)^2 = 1 + 3A(x)$.

◇

Exercise 2.8 Let $A \subset \mathbf{R}$ be non-empty. For $c \in \mathbf{R}$, put $c + A = \{c + a \mid a \in A\}$ and $cA = \{ca \mid a \in A\}$.

(a) Prove that $\inf A \leq \sup A$. What can you say if these numbers are equal?

(b) If $A \subset B$, then $\inf B \leq \inf A \leq \sup A \leq \sup B$.

(c) Prove that $\sup(c + A) = c + \sup A$.

(d) Find an expression for $\sup(cA)$; the answer will depend on whether $c > 0$, $c < 0$, or $c = 0$.

Illustrate each part with a sketch. ◇

Exercise 2.9 Let A and B be non-empty sets of real numbers. Prove that

$$\sup(A \cup B) = \max(\sup A, \sup B), \qquad \inf(A \cup B) = \min(\inf A, \inf B).$$

Suggestion: Either show that the right-hand sides satisfy the condition of being the sup/inf, or prove that

$$\sup A,\ \sup B \leq \sup(A \cup B) \leq \max(\sup A, \sup B).$$

Show that there can be no such formulas for the sup and inf of an *intersection* of sets. Make this principle as precise as you can. ◇

Exercise 2.10 For each family of intervals, prove that the union or intersection is as stated.

(a) $\displaystyle\bigcup_{n=1}^{\infty} [-n, n] = \mathbf{R}$.

(b) $\displaystyle\bigcap_{n=1}^{\infty} [0, 1 + 1/n) = [0, 1]$.

(c) $\bigcap_{n=1}^{\infty}(0, 1/n] = \varnothing$.

◇

Exercise 2.11 Complete the proof of Proposition 2.22 by showing that the set P defined by (2.11) satisfies the three order axioms. ◇

Exercise 2.12 Prove Proposition 2.31. ◇

Annuities and Amortization

Exercise 2.13 Commerce has been a driving force behind mathematical discovery since Babylonian times. In this exercise you will find a useful financial formula essentially from scratch.

When money is loaned, the lender usually charges the borrower a fee (called *interest*) that is proportional to the amount owed. Typically, the borrower pays the lender in installments of a fixed size at fixed time intervals (monthly or yearly, say). Part of each installment goes towards paying off the accrued interest, and part goes toward reducing the amount borrowed (the *principal*). The problem is to determine the size of each payment, given the amount borrowed, the interest rate, the time required to pay off the loan, and the number of payments.

(a) Let r be the annual interest rate in percent (you may assume $0 < r < 100$),[7] and suppose $n \geq 1$ payments are made each year. The interest rate per period is $r/n\%$, so the amount of interest accrued in a given period is $\rho := r/(100n)$ times the amount owed at the start of the period. After interest is added, that period's payment is subtracted, giving the new amount owed.

Let A_{i-1} be the amount owed at the start of the ith period, and let P be the payment. Find A_i in terms of A_{i-1}, ρ, and P. (Use the recipe at the end of the previous paragraph to compute the interest, then subtract the payment.) What happens financially if $P = \rho\, A_{i-1}$? If $P < \rho\, A_{i-1}$?

(b) Let A_0 be the amount borrowed initially. Determine the amount A_i owed after i payments have been made. (The indexing is consistent with the first part.)
Hints: Your answer will depend on A_0, ρ, P, and i. You might start by calculating the first few A_i by hand until you recognize a pattern.

[7] This ensures that $\rho < 1$ below. In fact, charging a rate over about 30% is a crime, called *usury*, though there is no mathematical reason to bound r.

Then prove your pattern is correct using part (a) and induction on i. Equation (2.10) should be helpful in getting the answer into closed form. You are cautioned to compute at least A_1, A_2, and A_3; there are some likely-looking patterns that are wrong.

(c) The loan is to be paid off when $i = kn$; use this to find the payment P in terms of the amount borrowed A_0, the "interest per period" multiplier ρ, and the total number $N := kn$ of payments. Observe that the payment is proportional to the amount borrowed, while the dependence on the interest rate is more complicated.

Suppose \$10,000 is borrowed at 4%, to be paid off in five years with equal monthly payments. How large is each payment? How much money is paid back in total? ◇

The Binomial Theorem and Applications

Let $n \geq 0$ be an integer. The *factorial* of n, denoted $n!$, is defined recursively by

$$0! = 1, \qquad n! = n\,(n-1)! \quad \text{for } n \geq 1. \tag{2.17}$$

If $n \geq 1$, then $n!$ is the product of the positive whole numbers not exceeding n. The convention for 0! is justified by Exercise 2.15. The *double factorial* $n!!$, not to be confused with $(n!)!$, is defined by

$$0!! = 1!! = 1, \qquad n!! = n\,(n-2)!! \quad \text{for } n \geq 2. \tag{2.18}$$

For example, $6!! = 6 \cdot 4 \cdot 2$ and $7!! = 7 \cdot 5 \cdot 3 \cdot 1$.

Exercise 2.14 Show that $n! = n!!\,(n-1)!!$ and $(2m)!! = 2^m\,m!$ for all $m, n \in \mathbf{N}$, both informally and by mathematical induction. ◇

The next exercise introduces "binomial coefficients" and a few natural ways they arise. If k and n are non-negative integers with $0 \leq k \leq n$, then the *binomial coefficient* $\binom{n}{k}$, read "n choose k," is defined to be

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}; \tag{2.19}$$

observe that $\binom{n}{n} = \binom{n}{0} = 1$ for $n \geq 0$. By definition, $\binom{n}{k} = 0$ if $k < 0$ or $k > n$. Though it is not immediately obvious, the binomial coefficients are *integers*; in fact, they have a useful combinatorial interpretation, see part (b).

Exercise 2.15 (a) Show that if $0 \leq k < \ell \leq n/2$, then $\binom{n}{k} < \binom{n}{\ell}$. (This can be done easily from the definition in more than one way.)

Use the definition to show that

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1} \qquad \text{for} \quad 1 \le k \le n, \tag{2.20}$$

and use this observation to make a table of the binomial coefficients up to $n = 5$. If you write the coefficients for fixed n in a row, then the entries in the next row are the sums of adjacent entries, and the resulting diagram is called *Pascal's triangle*. To get you started, the first four rows are:

$$\begin{array}{rccccccccccc} & \cdots & & 0 & & 0 & & 1 & & 0 & & 0 & \cdots \\ n=0 & \cdots & 0 & & 0 & & 1 & & 1 & & 0 & & 0 & \cdots \\ 1 & \cdots & & 0 & & 1 & & 2 & & 1 & & 0 & \cdots \\ 2 & \cdots & 0 & & 1 & & 3 & & 3 & & 1 & & 0 & \cdots \end{array}$$

Equation (2.20) essentially characterizes the binomial coefficients; knowledge of $\binom{n}{k}$ for all $k \in \mathbf{Z}$ (and for some $n \ge 0$) uniquely determines $\binom{n+1}{k}$ for all $k \in \mathbf{Z}$. In particular, the binomial coefficients are integers, because $\binom{0}{k}$ is an integer for every k.

(b) Let $\underline{\mathbf{n}}$ be a finite set having exactly $n \ge 0$ elements; then $\underline{\mathbf{0}} = \varnothing$, and for definiteness say $\underline{\mathbf{n}} = \{1, \dots, n\}$ if $n \ge 1$. Define $B(n,k)$ to be the number of subsets of $\underline{\mathbf{n}}$ that have exactly k elements. Clearly $B(0,0) = 1$, while $B(n,k) = 0$ if $k < 0$ or $k > n$. By writing $\{1, \dots, n+1\} = \{1, \dots, n\} \cup \{n+1\}$, show that

$$B(n+1,k) = B(n,k) + B(n,k-1) \qquad \text{for } n \ge 0,\ k \ge 1.$$

Use this to prove that $B(n,k) = \binom{n}{k}$ for all integers k and n.

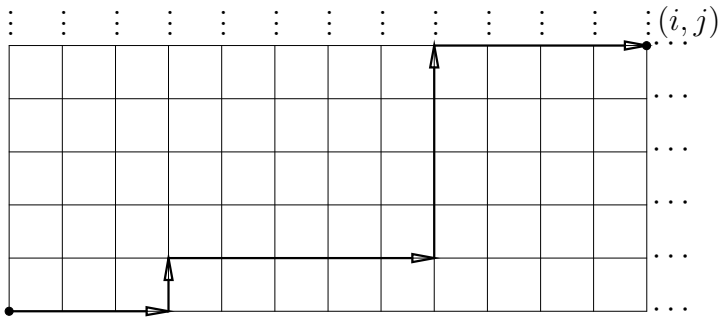
(c) An expression $(a+b)^n$ with a and b (real or complex) numbers and $n \in \mathbf{N}$ is a *binomial*. If this expression is multiplied out, the result will be a sum of terms of the form $a^k\, b^{n-k}$, since each term in the product has total degree n. Prove the *Binomial Theorem*:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k}. \tag{2.21}$$

Give as many arguments as you can; at least one should be a formal proof by induction, but there are other ways of bookkeeping, such as counting the number of times the monomial $a^k\, b^{n-k}$ appears in the product. Use the Binomial Theorem to prove that

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n, \qquad \sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0.$$

(d) Consider the unit square grid in the first quadrant.



Let $C(i, j)$ be the number of paths that join the origin to the point (i, j) by a sequence of upward and rightward steps along the grid. Argue that $C(i, j) = C(i-1, j) + C(i, j-1)$ for $i, j \in \mathbf{N}$, and that $C(i, 0) = 1$ for all $i \geq 0$.

Find a binomial coefficient that is equal to $C(i, j)$. (Suggestion: How many steps does it take to get from $(0, 0)$ to (i, j)? How many of these steps are horizontal/vertical? This should allow you to guess the answer; then you can verify the correctness of your guess, either by changing variables so the recursion relation for $C(i, j)$ becomes equation (2.20), or by induction on the number of steps.)

(e) Using a piece of graph paper, draw a "mod 2 Pascal's triangle" whose entries are 0 or 1 according to whether the corresponding binomial coefficient is even or odd. Filled/empty squares can be substituted for 0's and 1's. Try to include at least 32 rows. ◇

Representing Real Numbers as Decimals

Exercise 2.16 Let $0 < r < 1$, and let $a \in \mathbf{R}$. Equation (2.10) asserts that

$$\sum_{i=0}^{n} a\, r^i = a\, \frac{1 - r^{n+1}}{1 - r}.$$

Use Exercise 2.5 (c) to show that

$$\sup_{n \in \mathbf{N}} \sum_{i=0}^{n} a\, r^i = \frac{a}{1 - r} \qquad \text{for } 0 < r < 1. \tag{2.22}$$

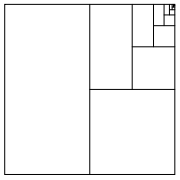
Evaluate this for $a = 0.9$ and $r = 0.1$. ◇

Intuitively, equation (2.22) gives a method for finding the "sum of infinitely many terms," provided these terms form a geometric progression. In fact, there is no "infinite addition"; a supremum is involved. Nonetheless, the standard notation is that

$$\sum_{i=0}^{\infty} ar^i = \frac{a}{1-r} \qquad \text{for } 0 < r < 1.$$

Setting $a = r = 1/2$ in equation (2.22) gives the suggestive result

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \cdots = 1,$$



Exercise 2.18 (c) gives a more substantial application.

Exercise 2.17 This exercise outlines the steps needed to fit decimal arithmetic into the axiomatic description of **R**. You should use only the axioms for **R** and **N**, though as always you are free to use all your knowledge to guess answers and strategies in advance.

In a decimal expression such as 314.1592, the location of a digit to the left or right of the decimal point determines a corresponding power of 10 by which the digit is multiplied. Specifically, if $a_1, \ldots, a_n$ and $b_0, b_1, \ldots, b_m$ are digits (that is, elements of the set $\{0, 1, 2, \ldots, 9\}$) then the expression $b_m \cdots b_1 b_0 . a_1 a_2 \cdots a_n$ stands for the (positive) rational number

$$\sum_{j=0}^{m} b_j \, 10^j + \sum_{i=1}^{n} a_i \, 10^{-i}.$$

The story is a bit more complicated if there are infinitely many nonzero a_i; this eventuality is treated in part (c).

(a) Prove that every real number x can be written, in exactly one way, as a sum $N + d$ for N an integer and $0 \le d < 1$.
Comments: It may help to consider the cases $x \ge 0$ and $x < 0$ separately. You will probably need the Archimedean property of **R** and Property **N**.3. The usual notation is $N =: \lfloor x \rfloor$ and $d = x \bmod 1$. When x is *positive*, N and d are called the *integer part* and *decimal part* of x; when $x < 0$, N and d are *not* the integer part and decimal part. For example, if $x = -3.14159$, then $N = -4$ and $d = 0.85841$, while the integer and decimal parts are -3 and -0.14159.

(b) Show that every natural number N can be written *uniquely* as

$$(*) \qquad b_m \cdots b_1 b_0 := \sum_{j=0}^{m} b_j \, 10^j, \qquad b_j \in \{0, 1, 2, \ldots, 9\}, \quad b_m \neq 0.$$

Comments: This may be so ingrained a truth that it will be difficult to decide what needs to be proven. Remember that the natural numbers are defined in terms of iterated successorship; there is nothing intrinsically "base 10" about $\mathbf{N}$. This question is intended more to get you thinking than to have you provide a complete proof, which could be lengthy depending on your level of skepticism. Here are some issues you should consider: Given a natural number N, is there *some* power of 10 larger than N? A largest power of 10 that is smaller than N? If so, how many times can this power be subtracted from N before the difference becomes negative? Does this reduce the quest to a simpler case? How do you determine which of two representations $(*)$ is larger?

(c) Show that every expression

$$(\dagger) \qquad 0.a_1 a_2 \cdots a_n := \sum_{i=1}^{n} a_i \, 10^{-i}, \qquad a_i \in \{0, 1, 2, \ldots, 9\}$$

is a rational number in the interval $[0, 1)$. If infinitely many of the a_i's are non-zero, then define

$$(\ddagger) \qquad \sum_{i=1}^{\infty} a_i \, 10^{-i} := \sup_{n \in \mathbf{N}} \sum_{i=0}^{n} a_i \, 10^{-i}.$$

Show that every expression $(\ddagger)$ represents a real number in $[0, 1]$. Prove, conversely, that every real number x with $0 \leq x \leq 1$ can be represented by an expression of the form $(\ddagger)$.
Comments: That the expression in $(\dagger)$ is smaller than 1 should be an easy induction on n, the number of digits. The only other subtle part is to show that every real number in $[0, 1]$ has a decimal expansion; you may not want to write out all the details, but should at least convince yourself it can be done. The idea is similar to that of part (b); it may be helpful to imagine the unit interval subdivided into tenths, hundredths, and so forth. A decimal representation of x can then be "read off" the location of x.

(d) Decimal representations in (‡) are not unique; for example, $1.0\overline{0} = 0.9\overline{9}$. Show that this is essentially the only ambiguity in the following sense. Two decimal expressions

$$\sum_{i=1}^{\infty} a_i\, 10^{-i} \quad \text{and} \quad \sum_{i=1}^{\infty} a'_i\, 10^{-i}$$

represent the same real number in $[0, 1)$ iff

- $a_i = a'_i$ for all $i \in \mathbf{N}$, or
- There is an $n \in \mathbf{N}$ such that $a_i = a'_i$ for $1 \le i < n$, $a_n = a'_n + 1 \le 9$, and $a_i = 0$, $a'_i = 9$ for $i > n$.

For example, $0.249\overline{9} = 0.25$. In the second case it may be necessary to reverse the roles of a_i and a'_i. ◇

Exercise 2.18 This continues Exercise 2.17, but can be done independently since the results of that exercise are familiar grade-school facts.

(a) Show that every rational number has a decimal representation that either terminates (is finite) or is eventually repeating (after finitely many decimal places, there is a finite string of digits that repeats *ad infinitum*, as in $1/12 = .083\overline{3}$). In fact, show that if p/q is in lowest terms and has eventually repeating decimal expansion, then the repeating string of digits is of length at most $q - 1$.
Hint: The decimal expansion of a rational number can be found by long division of q into p, and there are only finitely many possible remainders at each step of the process.

(b) Prove that every terminating or eventually repeating decimal represents a rational number.
Hints: This is clear for terminating decimals, see also Exercise 2.17 (c). For repeating decimals, it is enough to show that

$$0.a_1a_2\cdots a_N\overline{a_1a_2\cdots a_N}$$

represents a rational number (why?), and this can be accomplished with Exercise 2.16 (c), using $r = 10^{-N}$ and $a = .a_1a_2\cdots a_N \in \mathbf{Q}$.

(c) Write $.12\overline{12}$ and $.87544\overline{544}$ as rational numbers in lowest terms.
Comments: Part (c) is of course meant to ensure you understand part (b). In summary, this exercise shows that irrational numbers have

non-terminating, non-repeating decimal representations, and these are unique. Rational numbers whose repeating digits are not all "9" and not all "0" also have unique decimal representations. Terminating rationals have exactly two representations. ◇

Exercise 2.19 If $b \geq 2$ is a natural number, then everything done in Exercise 2.17 has an analogue using powers of b rather than powers of 10. The resulting notation is said to be "base b," though the special cases $b = 2$, 3, 8, and 16 are called binary, ternary, octal, and hexadecimal respectively. Formulate the analogous claims for each part of that exercise (particularly, what symbols are needed, and what rational number does a base-b expression stand for?), and convince yourself that their proofs are straightforward modifications of the arguments for decimal notation.

Write the decimal number 64 in ternary, octal, and hexadecimal notation. Write the fraction 1/3 in ternary and binary. ◇

Continued Fractions

For typographical simplicity, an expression

$$a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}} \tag{2.23}$$

with $a_k \in \mathbf{Z}$ for all k, $a_k > 0$ if $k > 1$, and $a_n > 1$ will be denoted $[a_1; a_2, \ldots, a_n]$. With these restrictions, the expression in (2.23) is called a *finite continued fraction*, and may be regarded as analogous to a finite decimal expression. The next exercises investigate the possibility of approximating real numbers by finite continued fractions, which leads to "infinite" continued fractions, analogous to infinite decimals. Continued fractions have a few theoretical advantages over decimals: the representation of $x \in \mathbf{R}$ is unique, is finite iff $x \in \mathbf{Q}$, and does not depend on a choice of base b. Continued fractions also provide, in a sense, "optimal" approximations to irrational numbers. The next two exercises are concerned with rational numbers and finite continued fractions; Exercise 2.22 treats irrational numbers and infinite continued fractions.

Exercise 2.20 Suppose throughout that $0 < q < p$, and that p and q have no common divisor.

(a) Set $r_0 = p$ and $r_1 = q$, and recursively define r_k for $k \geq 2$ by

$$r_{k-1} = a_k r_k + r_{k+1}, \qquad 0 \leq r_{k+1} < r_k. \tag{2.24}$$

(Cf. equation (2.25) below.) Define n to be the largest index for which $a_k \neq 0$. Prove that $a_k > 0$ for $1 \leq k \leq n$, and that

$$\frac{p}{q} = [a_1; a_2, a_3, \ldots, a_n].$$

Conclude that every rational number x has a unique finite continued fraction representation. (It is clear that every finite continued fraction represents a rational number.)

(b) Use part (a) to find the continued fraction representations of 5/7, 8/5, and 355/113.

(c) Express the continued fraction of q/p in terms of $[a_1; a_2, \ldots, a_n]$, the continued fraction of p/q.

(d) Does increasing a_n make $[a_1; a_2, a_3, \ldots, a_n]$ larger or smaller (or is the question more subtle than this, and if so, what's the real answer)?

If you get stuck on the last part, do the next exercise. ◇

Exercise 2.21 Fix a rational number $x = [a_1; a_2, \ldots, a_n]$, and for $1 \leq k \leq n$ define $r_k = [a_k; a_{k+1}, \ldots, a_n]$.

(a) Prove that $r_{k+1} = 1/(r_k - a_k)$ for $1 \leq k < n$. Cite an appropriate result from this chapter to conclude that if r_{k+1} is made larger, then r_k decreases, and *vice versa*.

(b) If $a_n = r_n$ is made larger, does $x = r_1$ increase or decrease?

There is nothing to this problem but elementary algebra, but the result will be very useful shortly. ◇

Exercise 2.22 Let $x \in \mathbf{R}$, and as in Exercise 2.17 let $\lfloor x \rfloor$ denote the greatest integer that is no larger than x. We wish to investigate the possibility of writing

$$x = [a_1; a_2, a_3, \ldots] = a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ddots}}$$

with a_k integers and $a_k > 0$ for $k \geq 1$. Briefly putting aside the question of what these infinite expressions mean, define for each positive integer n the numbers

$$x_n = [a_1; a_2, a_3, \ldots, a_n], \qquad r_n = [a_n; a_{n+1}, a_{n+2}, \ldots].$$

(a) Prove that $x_1 < x_3 < x_5 < \cdots < x_6 < x_4 < x_2$, namely that

$$x_{2k-1} < x_{2k+1} < x_{2k+2} < x_{2k} \qquad \text{for all } k \geq 1.$$

You should be able to read most of this off part (b) of the previous exercise. Show formally that $r_{k+1} = 1/(r_k - a_k)$ for $k \geq 1$, cf. Exercise 2.21 (a).

(b) Given $x \in \mathbf{R}$, recursively define integers a_k, p_k, and q_k, and real numbers y_k, as follows: Set $y_1 = x$, $p_{-1} = q_0 = 0$, and $p_0 = q_{-1} = 1$. Then define, for $k \geq 1$,

$$\begin{aligned} a_k &= \lfloor y_k \rfloor \\ p_k &= a_k\, p_{k-1} + p_{k-2} \\ q_k &= a_k\, q_{k-1} + q_{k-2} \\ y_{k+1} &= \frac{1}{y_k - a_k} \end{aligned} \tag{2.25}$$

Prove inductively that for all $n \geq 1$, $a_n > 0$, $q_n < q_{n+1}$ and $p_n < p_{n+1}$, $x_n = p_n/q_n$, and show formally that $y_n = r_n$.

(c) Prove inductively that $|x - x_n| < 1/(q_n q_{n+1})$ for $n \geq 1$. This gives a quantitative measure of the approximation of x by x_n.

(d) Prove that "continued fractions are the best rational approximations to x" in the following sense:

$$\left| x - \frac{p}{q} \right| < \frac{1}{q^2} \quad \text{iff} \quad \frac{p}{q} = \frac{p_n}{q_n} \quad \text{for some } n \in \mathbf{N}.$$

By part (a), the supremum of the numbers x_{2k+1} exists, and the infimum of the numbers x_{2k} exists. In fact, these bounds are equal, and their common value is *defined* to be the infinite continued fraction $[a_1; a_2, a_3, \ldots]$. We return to continued fractions in Chapter 4, when it will be easier to investigate these questions. ◇

Exercise 2.23 This exercise investigates the well-known relationship between "periodic" continued fractions and roots of quadratic polynomials.

(a) Let $x = \sqrt{2}$; calculate p_n and q_n for $1 \leq n \leq 6$. Find a pattern in the list a_1, a_2, $a_3, \ldots$ and prove your guess is correct. Calculate x_n^2 for $1 \leq n \leq 6$. You should not use an electronic calculator for any of this; it is enough to observe that $1 < \sqrt{2} < 1.5$. You will probably find it helpful to make a table like the following (with enough columns):

	$n = -1$	$n = 0$	$n = 1$	$n = 2$	$n = 3$	$\cdots$
a_n	$-$	$-$				$\cdots$
p_n	0	1				$\cdots$
q_n	1	0				$\cdots$
y_n	$-$	$-$				$\cdots$
x_n	$-$	$-$				$\cdots$

(b) Repeat for $x = \sqrt{3}$.

(c) Let a and b be positive integers. Show that the continued fraction $[a; b, a, b, a, \ldots]$ satisfies a quadratic polynomial with integer coefficients.

The integers appearing in the continued fraction of x can be used as a measure of "how irrational" x is. Rational numbers have finite continued fractions, while irrational roots of quadratic polynomials with rational coefficients have "periodic" continued fractions with period 2. $\diamond$

Chapter 3

Functions

The concept of "function" is absolutely central to all of modern mathematics, and in particular is one of the most important concepts in calculus. The colloquial sense of the word—"Your exam grade is a function of the effort you put into studying," or "Your standard of living is a function of income"—is similar to the mathematical one, and expresses a relation of *dependence of one thing upon another*. Functions are useful for modeling real-world relationships in mathematical terms.

Mathematically, a function may be regarded as a "rule" that assigns an "output" value (lying in some specified set) to each "input" (lying in another specified set); this is the common interpretation of a function as a "black box". Analysis is usually concerned with functions whose output is a real number and whose input is a finite list of real numbers, classically called "variables". The term "variable" is avoided in this book because it encourages denoting two entirely different concepts—numbers and functions—by the same symbol. In the author's view, "variables" are a *linguistic* concept rather than a mathematical one. Nonetheless, the term is sometimes convenient, and is used occasionally. In case of confusion, the definition is always the last word.

3.1 Basic Definitions

A function $f : X \to Y$ (read "f from X to Y") consists of three things:

- A non-empty set X, called the *domain* of f, whose elements are the "inputs" of f;
- A non-empty set Y, called the *range* of f, whose elements are the

"potential outputs" of f;

- A "rule" for assigning a unique element $y \in Y$ to each element $x \in X$. The element y is called the *value of f at x* and is often written $y = f(x)$.

These three pieces of information are conceptually analogous to the axioms for the natural numbers: They specify *properties* of functions that are used in practice. By contrast, the formal definition consists of an *implementation* of these properties using nothing but sets. This is our final definition at the level of sets; everything subsequent is defined using properties at the level of axioms.

Definition 3.1 Let X and Y be non-empty. A *function* $f : X \to Y$ is a subset $\Gamma_f = \Gamma \subset X \times Y$ such that for each $x \in X$, there is a unique $y \in Y$ with $(x, y) \in \Gamma$.

Many functions in this book are "real-valued" functions (meaning $Y \subset \mathbf{R}$) of a "real variable" (meaning $X \subset \mathbf{R}$). For these functions, the graph can be viewed as a subset of the Cartesian plane $\mathbf{R} \times \mathbf{R}$, Figure 3.1.

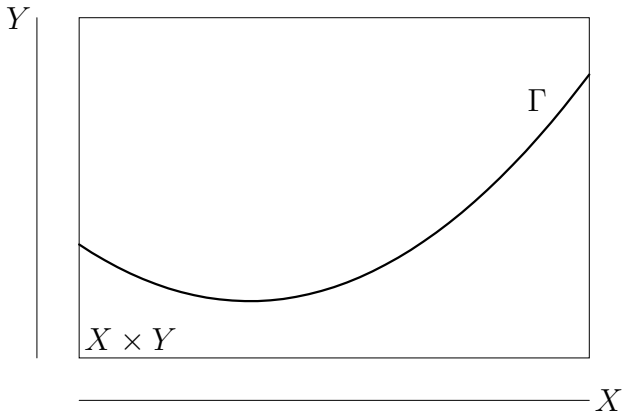


Figure 3.1: A function as a graph.

The greatest practical difference between this definition and the usual calculus book definition is that the domain and range are an essential part of f. Changing the domain—even by removing a single point—results in a *different function*. A function *does not* consist solely of a rule or formula; an equation like $f(x) = x^2$ does not define a function; see Example 3.2.

The set Γ is called the *graph* of f. By our definition, a function *is* its graph. We usually speak of "a function f" that "has graph Γ", though according to the definition we might well say "the function Γ".

Elements of the domain are called *points* and are said to be *mapped to* elements of Y by f; this is sometimes denoted $f : x \mapsto y$ or $x \mapsto y = f(x)$. Procedurally, begin with x, and find the unique $y \in Y$ such that $(x, y) \in \Gamma$, Figure 3.2.

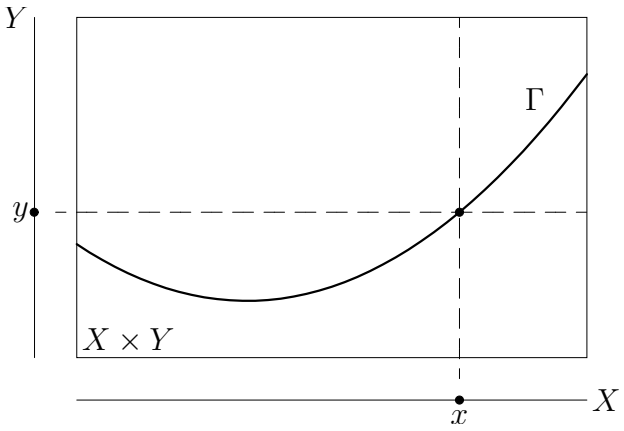


Figure 3.2: A function as a mapping.

We use the terms "function" and "mapping" to denote the same concept, though "function" suggests something real-valued while "mapping" does not.

Graphs and Procedures

Functions may be regarded "statically" or "dynamically." The graph of a function (statically) captures all features of the function in a single picture. Sometimes it is preferable to regard a function as a black box, so that $x \in X$ is a "variable" and the output $y = f(x)$ changes (dynamically) as x varies. In this picture, each x is a potential input, but the set of all inputs is not considered simultaneously.

For a physical example, consider a particle moving on a vertical number line, Figure 3.3. The domain is the set of t for which the motion is being considered, and the range is taken to be the entire real number line $\mathbf{R}$. In the dynamic picture, individual time values are taken, and the particle is regarded as moving up or down as t increases. In the static picture, the "history" of the particle is its world line, which

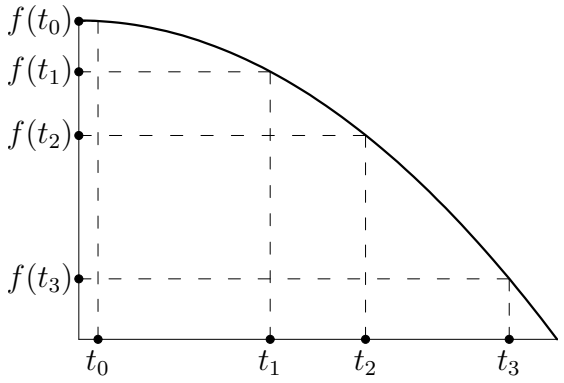


Figure 3.3: Static (bold) and dynamic interpretations of a function.

is exactly the graph of f. It is important to remember that these are two different views of the same mathematical situation, though it is often more convenient to look at a specific problem in one way or the other.

Surjectivity and Injectivity

A function has, by fiat, a unique value for every point x in the domain X. However, nothing in the definition asserts that every point y in the range Y *is* actually a function value. The set of all values of a function is its *image*, $f(X) = \{y \in Y : y = f(x) \text{ for some } x \in X\} \subset Y$:

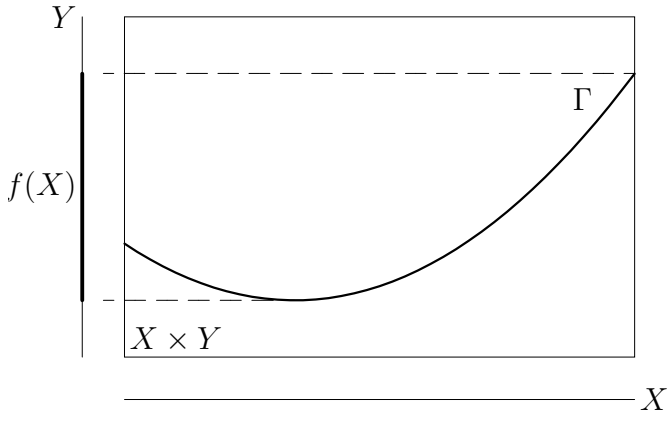


Figure 3.4: The image of a function.

A function f maps X *onto* Y if $Y = f(X)$, that is, if the image is the entire range. "Onto" is used as an adjective ("f is onto"), though the term *surjective*—coined by N. Bourbaki[1]—has the same meaning. The function depicted in Figure 3.4 is *not* surjective.

Most calculus books define the "range" of a function to be what we call the "image"; as a result, surjectivity is a superfluous concept. When working with a single function, it is often harmless to set the range equal to the image. However, when dealing with several functions whose images differ, it is important to distinguish the range from the image.

Injectivity

Though each x determines a unique y, nothing in the definition guarantees that each y in the image corresponds to a unique x; more than one point in the domain may be mapped to the same $y \in Y$, see Figure 3.5.

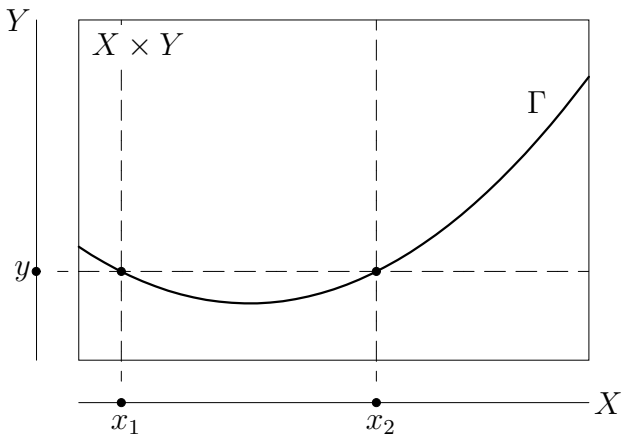


Figure 3.5: Distinct points can map to the same point in the image.

If every y in the image is the value of f for a *unique* $x \in X$, then the function f is *one-to-one*, or *injective*. In other words, injectivity means that "if $f(x_1) = f(x_2)$, then $x_1 = x_2$". This is also useful in the contrapositive form "if $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$," or in the form "it does not happen that $x_1 \neq x_2$ and $f(x_1) = f(x_2)$".

[1]Boor bah KEY: The pen name of an influential group of French mathematicians.

The Vertical and Horizontal Line Tests

Suppose $f : [a, b] \to [c, d]$, and let $R = [a, b] \times [c, d]$ be the rectangle in the x-y plane determined by the inequalities $a \leq x \leq b$ and $c \leq y \leq d$. The graph of f is a subset $\Gamma \subset R$ that "passes the vertical line test", namely, that intersects each vertical line $x = x_0$ (with $a \leq x_0 \leq b$) *exactly once*. Indeed, a vertical line meets the graph *at least once* because to every point of the domain is associated a function value, while the line intersects the graph *at most once* because a function is *single-valued*. These properties are guaranteed by the third clause above.

The conditions of injectivity and surjectivity have analogous geometric interpretations involving *horizontal* lines. Suppose $\Gamma \subset R$ is the graph of a function f. Then f is onto iff each line $y = y_0$ (with $c \leq y_0 \leq d$) intersects Γ *at least once*, while f is one-to-one iff every horizontal line intersects the graph *at most once*.

A function that is both one-to-one and onto is a *bijection*. The remarks above imply that the graph of a bijection $f : [a, b] \to [c, d]$ intersects each line $y = y_0$ ($c \leq y_0 \leq d$) *exactly once*; we might say the graph "passes the horizontal line test". Remember that whether or not a function is injective or surjective depends not only on the "rule" defining the function, but also on the domain and range.

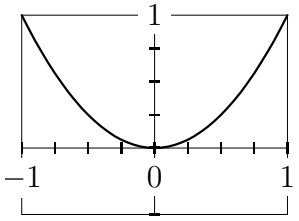


Figure 3.6: Functions associated to the squaring rule.

Example 3.2 Here are four different functions, all "defined" by the squaring rule $f : x \mapsto x^2$, but having different domains and/or ranges. Each of the functions below is obtained by excising a (possibly empty) portion of Figure 3.6.

- $X = [-1, 1]$ and $Y = [-1, 1]$. This function is not injective, since for example $-1 \neq 1$ but $f(-1) = f(1)$. Neither is the function surjective because, for example, $y = -1$ is not the square of a real number, hence is not in the image.

- $X = [-1, 1]$ and $Y = [0, 1]$; the bottom half of the figure is removed. This function is onto, because every real number $y \in [0, 1]$ *is* the square of some real number $x \in [-1, 1]$ by Theorem 5.8. As above, this function is not injective.

- $X = [0, 1]$, $Y = [-1, 1]$; the left half is removed. This function is not onto, as in the first example. However, this function *is* one-to-one, because the two square roots of a positive real number are negatives of each other, and only one of them is in the domain of f. Formally, if $f(x_1) = f(x_2)$, then

$$0 = f(x_1) - f(x_2) = x_1^2 - x_2^2 = (x_1 - x_2)(x_1 + x_2).$$

  Now, if x_1 and x_2 are points in the domain of f, then $x_1 + x_2 > 0$, so the previous equation implies $x_1 = x_2$. Thus f is injective.

- $X = [0, 1]$, $Y = [0, 1]$; only the upper right quadrant remains. This function is a bijection, as is easily checked from assertions above.

To emphasize one last time, changing the domain (and/or range) yields a *different function.* □

Monotone Functions

Let X be a set of real numbers, such as an interval. A function $f : X \to \mathbf{R}$ is said to be *increasing* if for all x_1 and x_2 in the domain, $x_1 < x_2$ implies $f(x_1) < f(x_2)$. Geometrically, the graph "slopes upward to the right". If n is a positive integer, then the nth power function $f : [0, +\infty) \to \mathbf{R}$ defined by $f(x) = x^n$, is increasing by Theorem 2.23. Similarly, a function $f : X \to \mathbf{R}$ is *decreasing* if $x_1 < x_2$ implies $f(x_1) > f(x_2)$. A function that is either increasing or decreasing is *strictly monotone.*

Application of an increasing function preserves inequalities. For example, the squaring function is increasing on the set of positive reals and $(1.7)^2 = 2.89 < 3 < 3.24 = (1.8)^2$, so if there exists a positive real number $\sqrt{3}$ whose square is 3, then $1.7 < \sqrt{3} < 1.8$, see Figure 3.7. Note that a strictly monotone function is injective. (Prove this from the definition if it's not obvious!) Similarly, application of a decreasing function reverses inequalities. Theorem 2.23 says that the reciprocal function is decreasing on the set of positive real numbers: If $0 < x_1 < x_2$, then $1/x_1 > 1/x_2$.

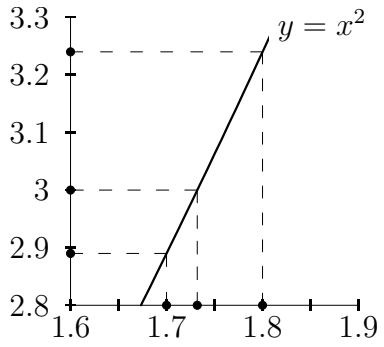


Figure 3.7: An increasing function preserves order relations.

A function f is *non-decreasing* if $x_1 < x_2$ implies $f(x_1) \leq f(x_2)$. An increasing function is certainly non-decreasing, but not conversely. For example, a constant function is non-decreasing, but not increasing. You should have no trouble writing down a definition of a *non-increasing* function. A function that is either non-decreasing or non-increasing is *monotone*.

Preimages

Let $f : X \to Y$ be a function. If $B \subset Y$, then the *preimage* of B under f is the set of all points of X that are mapped into B by f:

$$f^{[-1]}(B) = \{x \in X \mid f(x) \in B\} \subset X, \tag{3.1}$$

see Figure 3.8. Preimages satisfy some useful, easily-verified properties:

Proposition 3.3. *Let $f : X \to Y$ be a function. If $A \subset X$ and $B \subset Y$, then*

$$f\big(f^{[-1]}(B)\big) = B \quad \text{and} \quad A \subset f^{[-1]}\big(f(A)\big).$$

The second inclusion is generally proper.

Proof. Exercises 3.2 and 3.3. □

The preimage of B may be empty even if B is non-empty (if f is not onto), and the preimage of a one-point set may consist of more than one point (if f is not one-to-one). Consequently, if f is not bijective there is no way to regard $f^{[-1]}$ as a mapping from Y to X.

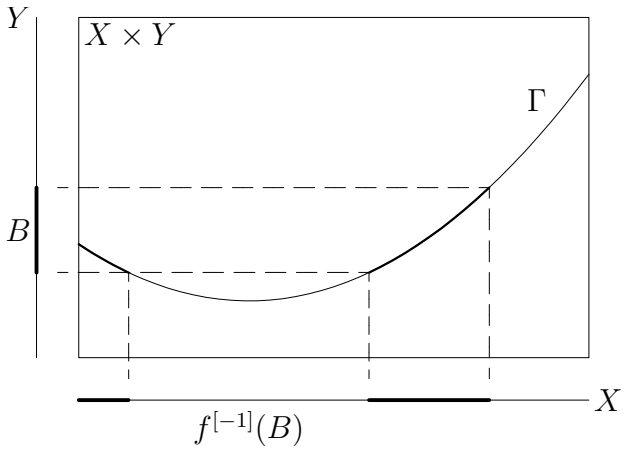


Figure 3.8: The preimage of a set B under f.

Restriction and Extension

One of the simplest things that can be done to a function is to make its domain smaller without changing the range; remember that this gives a *different function*. If $f : X \to Y$ is a function and $A \subset X$ is non-empty, then the *restriction* of f to A, denoted $f|_A$, is formally defined as $\Gamma \cap (A \times Y)$, see Figure 3.9. Loosely, the restriction is "given by the same rule as f, but is only defined for points of A." Said yet another way, $f|_A : A \to Y$ is the function defined by

$$f|_A(x) = f(x) \quad \text{for } x \in A. \tag{3.2}$$

Restriction therefore amounts to forgetting about part of f, or throwing away information. See Exercise 3.6 for relationships between injectivity, surjectivity, and restriction.

Extension

Let $f : A \to \mathbf{R}$ be a function and $B \supset A$. An *extension* of f to B is a function $F : B \to \mathbf{R}$ that agrees with f on A, that is, with $F|_A = f$. Extensions are never unique, but in real problems there are usually additional constraints, subject to which an extension may be unique or may not exist at all.

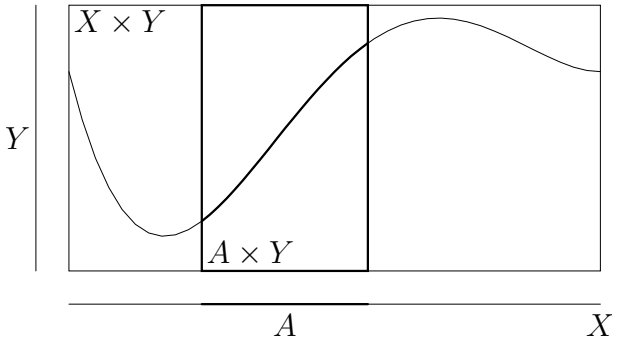


Figure 3.9: Restriction of a function.

Notational Cautions

When a function is defined by a formula such as

$$f(x) = \frac{x+1}{x-1}, \quad g(z) = \sqrt{1-z^2}, \quad \text{or } h(t) = \log t,$$

and the domain and range have not been specified explicitly, it is customary to agree that the range is the set $\mathbf{R}$ of real numbers, while the domain is the "natural domain," consisting of all real numbers for which the expression is a real number. When working with functions of a complex variable, this laxity is often inadequate; the domain must be specified carefully. In any case, it is better to be slightly over-careful than ambiguous.

It is badly ambiguous to talk about "the function x^2" for two reasons:

- The domain and range have not been specified (less serious);
- There is no mention of "an independent variable" (a grave omission).

Failing to declare the name of the independent variable is an extremely common error, because many students have learned by osmosis that x always stands for the "independent variable". The reverse point was carried to humorous extreme by a graffito in the Berkeley math building:

$\sqrt{3} > 2$ for sufficiently large 3.

We are so used to the symbol "3" denoting a particular real number that it is ridiculous to ascribe it another meaning. Unfortunately, the

convention of letting x denote the "independent variable" cannot reasonably be followed universally, as we shall see in later chapters. The rule $x \mapsto x^2$ is quite different from the rule $t \mapsto x^2$, yet there are real-life situations where one wants to consider a constant function of t whose value is everywhere x^2. The rule $x \mapsto x^2$ is a *procedure*, namely to take a number and square it, while x^2 is merely a number, albeit the one that happens to arise as output when x is the input. It would be more accurate to denote the "squaring rule" by

$$\square \longmapsto \square^2,$$

with the understanding that an arbitrary expression representing a number can be put into the box. This notation avoids welding the letter x to a specified role, but is too unwieldy for general use. In any case, the rules $x \mapsto x^2$, $t \mapsto t^2$, and $\xi \mapsto \xi^2$ are mathematically *identical* in the absence of further knowledge about the letters x, t, or ξ. Though x often denotes a "generic" input and y denotes the corresponding function value, it is equally permissible to reverse their roles, or to use some more complicated expression as the input. It is the *relationships* between input and output values that a function specifies, not the *notation*. The distinction between notation and meaning is one of the most difficult psychological points to absorb about mathematics.

The common construction "$x = x(t)$" should be used with extreme care, or (better) avoided altogether. On the left, x is a number; on the right, x is a function. Calling x a "variable" leads to murky syntax: Is x a function or a number? Is "$f(x)$" a function value or a 'composite function'? The problem is compounded the more functions are present, and can easily result in two different functions being assigned the same name. In the best of circumstances this causes needless confusion, but if it happens while you are using a symbolic manipulation program, you will come to grief: A computer program cannot distinguish objects except by their literal name.

Functions are "Black Boxes"

A function is completely specified by its domain, range, and the values it takes on points of the domain. While this statement sounds vacuous as an abstraction, it can be counterintuitive in practice. For example, each of these three formulas defines the absolute value function on $\mathbf{R}$:

$|x| = \sqrt{x^2}$;

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases} \qquad |x| = \begin{cases} x^2 & \text{if } x = -1,\ 0,\ \text{or } 1 \\ \sqrt{x^2} & \text{otherwise} \end{cases}$$

It is easy to find infinitely many other collections of formulas that define exactly the same function. None of them is more correct than the others, though the second of them happens to be the *definition*. This example is a little silly (though note the first formula, an identity that is often misremembered) because at this stage we have very few ways of defining functions other than as collections of algebraic formulas, so verifying equality of two functions is likely to be a matter of algebra. In subsequent chapters, functions with complicated definitions (involving limits and suprema, such as derivatives, definite integrals, and infinite sums) are studied; the so-called "Fundamental Theorem of Calculus" asserts that certain pairs of functions are equal. It often happens that one function has an interesting interpretation (but a complicated definition) while the other is easy to calculate (but has no interesting interpretation). Knowledge that these functions are equal is valuable information. To emphasize: "Equality" of functions f and g means that f and g have the same domain and the same range, and that $f(x) = g(x)$ for all x in the domain; the values $f(x)$ and $g(x)$ may be arrived at by completely different means.

3.2 Basic Classes of Functions

This section presents several interesting classes of functions. Some of them (such as polynomials and vectors) may be familiar; if so, you should try to mesh your existing knowledge with the framework of sets and functions explained so far.

Polynomials and Rational Functions

A *polynomial function* on a non-empty set $A \subset \mathbf{R}$ is a function $p : A \to \mathbf{R}$ defined by a formula

$$p(x) = \sum_{k=0}^{n} a_k\, x^k = a_0 + a_1\, x + a_2\, x^2 + \cdots + a_n\, x^n \tag{3.3}$$

with $a_k \in \mathbf{R}$ for $k = 0, \ldots, n$. Polynomial functions are important for many reasons. Not least is the fact that a polynomial function is evaluated using nothing but addition and multiplication.

The expression on the right-hand side of (3.3) is called a "polynomial (in x)," and should be distinguished from a polynomial *function* (which has a specified domain A). The number a_k is called the *coefficient* of x^k; the coefficient a_0 is often called the *constant term*. In the representation (3.3), it is usually assumed that $a_n \neq 0$, in which case the polynomial p is said to have *degree* n and a_n is called the *top degree* or *leading* coefficient. A polynomial is *monic* if the leading coefficient is 1.

Lemma 3.4. *Let $p : \mathbf{R} \to \mathbf{R}$ be the polynomial function given by (3.3). If $p(x) = 0$ for all x, then $a_k = 0$ for all k.*

Proof. We will do induction on $\deg p$, the degree of p. The result is obvious if $p(x) = a_0$ is a constant polynomial, i.e., if $\deg p = 0$. Suppose the conclusion of the lemma holds for every polynomial of degree n, and let p be of degree $(n+1)$. Write

$$p(x) = \sum_{k=0}^{n+1} a_k x^k = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n+1} x^{n+1}.$$

By hypothesis, $0 = p(0) = a_0$, so we have

$$0 = p(x) = \sum_{k=1}^{n+1} a_k x^k = x(a_1 + a_2 x + \cdots + a_{n+1} x^n) =: x q(x)$$

for all x; thus $q(x) = 0$ except possibly when $x = 0$. If we show that $q(0) = 0$, then we can apply the inductive hypothesis to conclude that the remaining coefficients of p are all zero.

We prove the contrapositive: If $a_1 = q(0) \neq 0$, then $q(x) \neq 0$ for some x. The idea is that for $|x|$ "very small", the value $q(x)$ is "approximately" a_1. Precisely, the reverse triangle inequality asserts that

$$(*) \quad |q(x)| = |a_1 + a_2 x + \cdots + a_{n+1} x^n| \geq \big| |a_1| - |x| \cdot |a_2 + \cdots + a_{n+1} x^{n-1}| \big|.$$

To get a lower bound, we seek an upper bound on $|x| \cdot |a_2 + \cdots + a_{n+1} x^{n-1}|$. The triangle inequality implies

$$(**) \qquad |x| \cdot |a_2 + \cdots + a_{n+1} x^{n-1}| \leq |x| \cdot \big(|a_2| + \cdots + |a_{n+1}|\, |x|^{n-1} \big).$$

If we pick x such that $|x| < 1$, then the right-hand side of $(**)$ is no larger than $|x| \cdot \big(|a_2| + \cdots + |a_{n+1}|\big)$. If in addition we take $|x| < |a_1|/2\big(|a_2| + \cdots + |a_{n+1}|\big)$, then the right-hand side of $(**)$ is no larger than $\frac{1}{2}|a_1|$. This, in turn, implies that the right side of $(*)$ is *at least* $\frac{1}{2}|a_1|$.

To summarize, we have shown that if $|a_1| > 0$, then

$$0 < |x| < \min\Big(1, \frac{|a_1|}{2\big(|a_2| + \cdots + |a_{n+1}|\big)}\Big) \implies |q(x)| \geq \frac{|a_1|}{2} > 0.$$

This is the desired contrapositive. □

For each fixed number a, a polynomial can be written in powers of $x - a$. You can think of x and $u := x - a$ as being two different coordinate systems; writing a polynomial in powers of $x - a$ describes the same polynomial to an observer in the "u world". For example, $1 + x^2 = 5 + 4u + u^2$ if $u = x - 2$. The general formula

$$\sum_{k=0}^{n} a_k\, x^k = \sum_{k=0}^{n} b_k\, (x - a)^k$$

determines the b_k in terms of the a_k by expanding and equating like powers of x. The polynomial representation on the right is said to be *in powers of* $x - a$. (Note that k is a dummy index, so the individual terms on the two sides of this equation are not equal; the summation signs cannot be dropped.) The top degree coefficients are always equal: $a_n = b_n$. In Chapter 14 we will obtain a fast calculational procedure for finding the b_k in terms of the a_k, see Example 14.12.

Arithmetic Operators as Functions

The domain of a function need not be a set of real numbers. Addition and multiplication can be viewed as real-valued functions whose domain is the set $\mathbf{R} \times \mathbf{R}$ of *ordered pairs* of real numbers; s and $p : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ (for "sum" and "product") are defined by

$$s(x, y) = x + y, \qquad p(x, y) = x \cdot y.$$

The systematic study of functions of "several variables" is undertaken in more advanced courses. It is easy to define the concept of "functions of several variables," but developing calculus for them is more difficult, and requires a solid understanding of calculus in one variable.

Lagrange Interpolation Polynomials

A common question on intelligence tests is to give the first few terms of a finite sequence—such as 1, 2, 3, 5, or $\otimes \oplus \oplus \otimes \otimes\otimes$—and to ask for the next term, or for the rule that generates the sequence. Ironically, such questions are non-mathematical, because no matter what pattern is given, there are infinitely many ways of continuing. These tests *do* demonstrate the remarkable ability of the human brain to discern patterns, even when no pattern is logically implied.

Suppose we wish to find a polynomial p that produces the numerical sequence above, in the sense that

$$(*) \qquad p(1) = 1, \quad p(2) = 2, \quad p(3) = 3, \quad p(4) = 5.$$

While there are infinitely many polynomials that satisfy these four equations, there is a *unique* such polynomial of degree 3 or less. This is not exactly obvious, but can be seen as follows. Imagine first that we had at our disposal four cubic polynomials e_1, e_2, e_3, and e_4 satisfying

$$\begin{array}{llll} e_1(1) = 1 & e_1(2) = 0 & e_1(3) = 0 & e_1(4) = 0 \\ e_2(1) = 0 & e_2(2) = 1 & e_2(3) = 0 & e_2(4) = 0 \\ e_3(1) = 0 & e_3(2) = 0 & e_3(3) = 1 & e_3(4) = 0 \\ e_4(1) = 0 & e_4(2) = 0 & e_4(3) = 0 & e_4(4) = 1 \end{array}$$

Then $p(x) = e_1(x)+2e_2(x)+3e_3(x)+5e_4(x)$ would be a cubic polynomial satisfying $(*)$, since for example (reading down the third column)

$$\begin{aligned} p(3) &= e_1(3) + 2e_2(3) + 3e_3(3) + 5e_4(3) \\ &= 0 + (2 \cdot 0) + (3 \cdot 1) + (5 \cdot 0) = 3. \end{aligned}$$

In fact, given the "magic polynomials" $\{e_i\}_{i=1}^4$ we could generate an arbitrary sequence of four numbers, just by filling in the blanks:

$$(**) \qquad p(x) = \_\_\_e_1(x) + \_\_\_e_2(x) + \_\_\_e_3(x) + \_\_\_e_4(x).$$

Now, how could we find the e_i? This is easier than it looks; the polynomial $\widetilde{e}_1(x) = (x-2)(x-3)(x-4)$ is non-zero at 1, and is zero at the other three numbers. Dividing by $(1-2)(1-3)(1-4) = -6$ adjusts the value at 1 and gives e_1:

$$e_1(x) = \frac{(x-2)(x-3)(x-4)}{(1-2)(1-3)(1-4)} = -\frac{x^3 - 9x^2 + 26x - 24}{6},$$

see Figure 3.10. Analogous reasoning tells us that

$$e_2(x) = \frac{(x-1)(x-3)(x-4)}{(2-1)(2-3)(2-4)},$$

$$e_3(x) = \frac{(x-1)(x-2)(x-4)}{(3-1)(3-2)(3-4)},$$

$$e_4(x) = \frac{(x-1)(x-2)(x-3)}{(4-1)(4-2)(4-3)}.$$

Of course, these polynomials can be multiplied out, but the form given makes it clear that these are the sought-after magic polynomials.

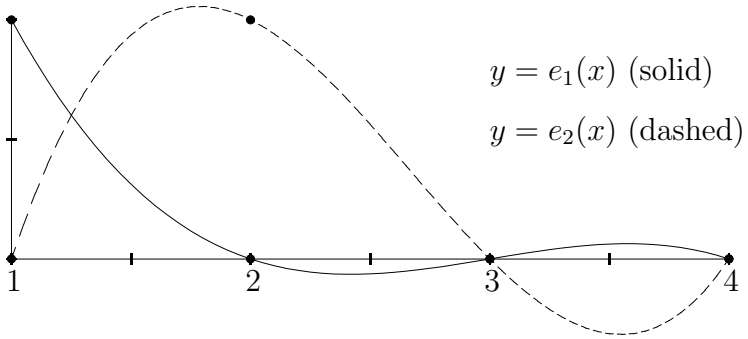


Figure 3.10: Interpolating four points with a cubic polynomial.

Once you have digested this solution, you will realize that the argument proves much more:

Theorem 3.5. *Let $B = \{b_i\}_{i=1}^n$ be a set of n* distinct *real or complex numbers, and let $C = \{c_i\}_{i=1}^n$ be an arbitrary set of n numbers. There exists a unique polynomial p of degree at most $n-1$ such that*

$$p(b_i) = c_i \qquad \text{for all } i = 1, \ldots, n. \tag{3.4}$$

The polynomial p whose existence is asserted by Theorem 3.5 is called the *Lagrange interpolation polynomial* for the given data $\{b_i\}$ and $\{c_i\}$. Existence of interpolating polynomials proves that every finite sequence of numbers can be generated by a polynomial of sufficiently large degree.

Proof. You should have little trouble convincing yourself that the key is to find a set of n "magic polynomials" $\{e_i\}_{i=1}^n$, each of which has degree $n-1$, is equal to 1 at b_i, and is zero at all the other b_j. This is a straightforward generalization of the argument above for four distinct points. The obvious generalization of $(**)$ allows you to interpolate an arbitrary sequence of n numbers.

The only new ingredient in the theorem is the uniqueness assertion. Suppose q is another polynomial of degree at most $n-1$ that satisfies (3.4). Then the difference, $p-q$, is a polynomial of degree at most $n-1$ that has n distinct roots, namely the b_i. This implies that $p-q$ is identically zero, so $p=q$. □

Piecewise Polynomial Functions

A function f on a closed, bounded interval is *piecewise polynomial* if the domain can be divided into a finite collection of intervals such that f is polynomial on each subinterval. An example is the function $f:[-2,2]\to\mathbf{R}$ defined by

$$f(x)=\begin{cases}0 & \text{if } -2\le x<-1\\ x^2 & \text{if } -1\le x\le 1/4\\ x-x^5 & \text{if } 1/4<x\le 1\\ 1/2 & \text{if } 1<x\le 2\end{cases}$$

The graph is depicted in Figure 3.11, which also illustrates the use of circles and spots to denote open or closed intervals.

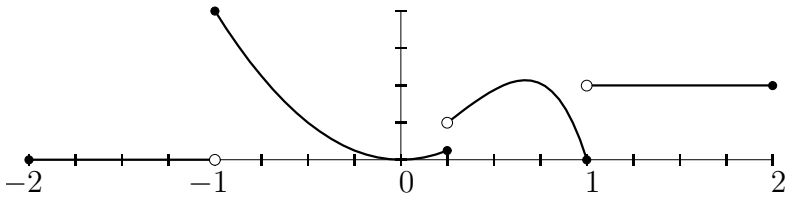


Figure 3.11: A piecewise-polynomial function.

Formal Power Series

A polynomial may be regarded as an expression $p(x)=\sum_{k=0}^{\infty}a_k\,x^k$ in which all but finitely many of the coefficients a_k are zero. The degree

is the largest index for which the corresponding coefficient is non-zero. With this notation, the sum and product of two polynomials is

$$\begin{aligned} \sum_{k=0}^{\infty} a_k\, x^k + \sum_{k=0}^{\infty} b_k\, x^k &= \sum_{k=0}^{\infty} (a_k + b_k)\, x^k, \\ \Big(\sum_{k=0}^{\infty} a_k\, x^k\Big) \cdot \Big(\sum_{k=0}^{\infty} b_k\, x^k\Big) &= \sum_{k=0}^{\infty} \Big(\sum_{i=0}^{k} a_i\, b_{k-i}\Big) x^k. \end{aligned} \tag{3.5}$$

The meaning of the first equation should be clear, while the second says that to multiply two polynomials, we multiply every summand of the first by every summand of the second, then gather like powers of x: The coefficient of x^k is

$$\sum_{i=0}^{k} a_i\, b_{k-i} = a_0 b_k + a_1 b_{k-1} + a_2 b_{k-2} + \cdots + a_{k-1} b_1 + a_k b_0.$$

In particular, the sum or product of polynomials is a polynomial. These equations have an obvious analogue for polynomials centered at a.

If we drop the assumption that at most finitely many coefficients are non-zero, then equation (3.5) still makes sense (calculation of each coefficient involves only finitely many arithmetic operations). Expressions of the form $p(x) = \sum_{k=0}^{\infty} a_k x^k$, called *formal power series*, can be added and multiplied unambiguously. A formal power series does not define a function of x in an obvious way, because it makes no sense to add infinitely many numbers together. Instead, a formal power series should be regarded as an infinite list $(a_k)_{k=0}^{\infty}$ of coefficients; equation (3.5) explains how to add and multiply two such lists. Formal power series are useful in many contexts. To give but one example, note that

$$(1 - x)(1 + x + x^2 + x^3 + \cdots) = 1.$$

Whether or not this equation has any meaning when x is assigned a specific numerical value is another matter entirely. It is sometimes possible to "add up" the terms of a formal power series (possibly with restrictions on x), thereby creating functions that are not polynomial. Such a function is said to be "(real) analytic". The functions studied before the Nineteenth Century—polynomials, rational functions, exponentials, logarithms, trigonometric functions, and a host of more exotic creatures encountered in classical analysis—are analytic. Mathematicians of the day tacitly assumed "functions" were analytic. The most

famous and prolific mathematician of the Eighteenth Century, L. Euler[2] was a profound genius at manipulating power series. Many of the spectacular results we encounter later in the book are due to Euler.

Polynomial Division, and Factorization

Let $\mathbf{F}$ be a field. A polynomial p with coefficients in $\mathbf{F}$ is said to *factor over* $\mathbf{F}$ if there exist *non-constant* polynomials p_1 and p_2 with coefficients in $\mathbf{F}$ that satisfy $p_1 p_2 = p$. For example, over $\mathbf{R}$ we have

$$x^2 - 2 = (x - \sqrt{2})(x + \sqrt{2}), \qquad x^3 - 2x - 4 = (x - 2)(x^2 + 2x + 2),$$
$$x^4 - 1 = (x - 1)(x + 1)(x^2 + 1).$$

A polynomial without factors is *irreducible* (over $\mathbf{F}$). The first example is irreducible over $\mathbf{Q}$, while the quadratic $x^2 + 1 = (x + i)(x - i)$ is irreducible over $\mathbf{R}$ but factors over $\mathbf{C}$. Enlarging a field makes it "easier" to factor polynomials.

There is a polynomial division algorithm with remainder, analogous to integer division with remainder. The following special case is adequate for our purposes in Chapter 15. The general case is similar, see Exercise 3.12.

Theorem 3.6. *Let p be a polynomial with coefficients in a field $\mathbf{F}$, and let $a \in \mathbf{F}$. There exists a unique polynomial q with coefficients in $\mathbf{F}$ and of degree less than* $\deg p$ *such that*

$$p(x) = (x - a)q(x) + p(a).$$

Proof. It suffices to prove the theorem when p is monic, since we may absorb a multiplied constant in q. We proceed by induction on the degree of p. The theorem is obvious if p is constant: take $q = 0$.

The statement, "If p is monic and of degree k, then there exists a polynomial q of degree at most $k - 1$ such that $p(x) = (x - a)(q(x) + p(a)$," is our inductive hypothesis at level k. Suppose that p is a monic polynomial of degree $(k + 1)$; the polynomial $p(x) - (x - a)x^k$ is of degree at most k (since we have subtracted off the term of highest degree). After factoring out the leading coefficient, we may apply the inductive hypothesis to find a q of degree at most $(k - 1)$ such that $p(x) - (x - a)x^k = (x - a)q(x) + p(a)$, or

$$p(x) = (x - a)\big(x^k + q(x)\big) + p(a).$$

[2]Pronounced "Oiler".

This is the inductive hypothesis at level $k+1$. □

Corollary 3.7. *A polynomial p is evenly divisible by $(x-a)$ iff $p(a)=0$.*

A number $a \in \mathbf{F}$ is a *root* of p if $p(a)=0$. Corollary 3.7 says there is a correspondence between roots and linear factors.

Rational Functions

A quotient of polynomials determines a "rational function." More precisely, if $A \subset \mathbf{R}$, then a function $f : A \to \mathbf{R}$ is a *rational function* if there exist polynomials p and q such that $q(x) \neq 0$ for all $x \in A$, and $f(x) = p(x)/q(x)$ for all $x \in A$. The polynomial q can be made non-vanishing on A by restriction (if necessary). Usually it is assumed that p and q have no non-constant factor in common, in which case the fraction p/q is said to be *reduced*. The expression $(1-x)/(1-x^2)$ is not reduced, while $1/(1+x)$ and $(1+x^2)/(1-x^2)$ are reduced.

The natural domain of $f(x) = (1-x)/(1-x^2)$ is the set of real numbers for which the denominator does not vanish, namely $\mathbf{R} \setminus \{\pm 1\}$. In this example, we can cancel a common factor, obtaining the function $g(x) = 1/(1+x)$, whose natural domain omits only $x=-1$. Observe that $f(x) = g(x)$ for $x \neq 1$, but formally $f(1) = 0/0$, while $g(1) = 1/2$: Canceling the common factor allows us to assign a value to $f(1)$. However, not canceling allows implicit restriction of the domain, which is sometimes useful. When finding the natural domain of a rational function $f = p/q$, ask where $q(x) = 0$ *without* canceling common factors.

Implicit and Algebraic Functions

A function need not be given "explicitly" by a formula in one variable, but may be specified "implicitly" by a relation in two variables. For example, the equation $x^2 + y^2 - 1 = 0$ defines two functions, $f(x) = \pm\sqrt{1-x^2}$ for $|x| \leq 1$. If we set $y = f(x)$, then the implicit relation $x^2 + y^2 - 1 = 0$ is satisfied for all x in the domain of f.

Let $R = (a,b) \times (c,d)$ be a rectangle in the plane, and let $F : R \to \mathbf{R}$ be a function. We say that the equation $F(x,y) = 0$ defines an *implicit function* in R if there exists a *unique* function $f : (a,b) \to (c,d)$ such that

$$F(x,y) = 0 \quad \text{for } (x,y) \in R \iff y = f(x) \quad \text{for } x \in (a,b).$$

Several rectangles are depicted in Figure 3.12; each "good" rectangle determines an implicit function, and each "bad" rectangle fails to. To emphasize, whether or not an equation defines an implicit function depends not only on the equation, but on the rectangle R in the plane. It is not terribly important to take an open rectangle.

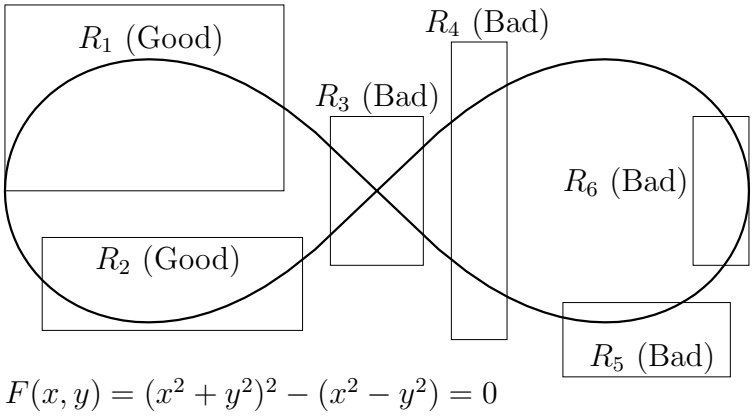


Figure 3.12: The zero locus of an algebraic equation, and implicit functions.

Example 3.8 The equation $x^2+y^2-1=0$ defines an implicit function in the rectangles $[-1,1]\times[0,1]$, $(0,1)\times(0,2)$, and $[-0.1,0)\times(-1.1,0.8)$, but not in the square $[-1,1]\times[-1,1]$, nor in the rectangle $[1-\delta,1]\times[-1,1]$, no matter how small $\delta>0$ is. You should draw a sketch and convince yourself of these assertions. □

Algebraic Functions

Let $F : \mathbf{R}\times\mathbf{R}\to\mathbf{R}$ be a polynomial function; this means that there exist constants a_{ij}, with $i,\, j\in\mathbf{N}$, and only finitely many a_{ij} non-zero, such that

$$F(x,y)=\sum_{i,j=0}^{\infty} a_{ij}\, x^i y^j \quad \text{for all } (x,y)\in\mathbf{R}\times\mathbf{R}. \tag{3.6}$$

The *zero locus* of F is $Z(F)=\{(x,y)\mid F(x,y)=0\}\subset\mathbf{R}\times\mathbf{R}$.

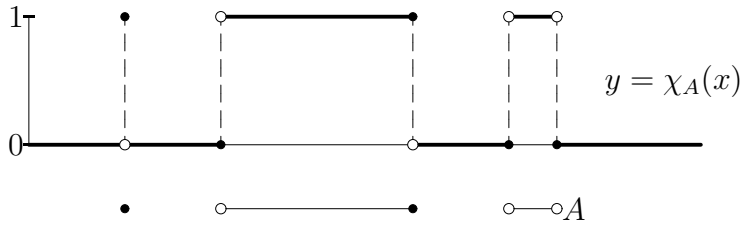
Definition 3.9 Let $f:(a,b)\to(c,d)$ be a function. If f is defined implicitly in the rectangle $(a,b)\times(c,d)$ by a *polynomial* function F, then we say f is an *algebraic function*.

Every rational function is algebraic (Exercise 3.8), but not conversely; as shown above, the function $f(x) = \sqrt{1-x^2}$ for $-1 < x < 1$, $0 \leq y < \infty$ is algebraic. Generally, it is *impossible* to express an algebraic function using only the four arithmetic operations and extraction of radicals, not merely as a practical matter, but as a theoretical one.

Characteristic and Step Functions

Let X be a non-empty set, and let $A \subseteq X$ (possibly empty). The *characteristic function* of A in X (sometimes called the "indicator function") is the function $\chi_A : X \to \mathbf{R}$ defined by

$$\chi_A(x) = \begin{cases} 1 \text{ if } x \in A, \\ 0 \text{ if } x \notin A. \end{cases} \tag{3.7}$$



The characteristic function answers—for each $x \in X$—the question, "Are you an element of A?" and converts the response into binary ("Yes" = 1, "No" = 0). Boolean operations are converted into arithmetic modulo 2, see Exercises 2.1 and 3.4. A computer scientist might take the range to be the finite field $\mathbf{F}_2 = \{0, 1\}$ rather than $\mathbf{R}$, to exploit the power of mod 2 arithmetic.

Let $I \subset \mathbf{R}$ be an interval. A *step function* is a function $f : I \to \mathbf{R}$ that takes only finitely many values, and whose preimages are all finite unions of points and intervals. As an example, for $k = 1, \ldots, n$, let I_k be an interval, χ_k the indicator function of I_k, and c_k a real number, and assume the intervals I_k are pairwise disjoint. The function

$$f(x) = \sum_{k=1}^{n} c_k \chi_k(x) = \begin{cases} c_k & \text{if } x \in I_k \\ 0 & \text{if } x \notin I_k \text{ for all } k \end{cases} \tag{3.8}$$

is a step function. In other words, a step function is not merely piecewise polynomial, but is "piecewise constant". Step functions are fundamental to the theory of integration (Chapter 7), both because they

can be "integrated" using only addition and multiplication, and because they can be used to approximate very large classes of functions. Exercise 3.5 characterizes step functions along the lines of (3.8).

Vectors and Sequences

The simplest functions are those whose domain is a finite set. The prototypical set with n elements is the *initial segment* $\underline{\mathbf{n}} := \{1, \ldots, n\}$. A function $\mathbf{x} : \underline{\mathbf{n}} \to \mathbf{R}$ is a collection of n ordered pairs, $\{(k, x_k) \mid 1 \leq k \leq n\}$. The same data is encoded as an *ordered n-tuple* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, often called a *vector*. Existence of Lagrange interpolation polynomials shows that every vector is a polynomial function; this observation is mostly a curiosity.

The *set of all real-valued functions* on $\underline{\mathbf{n}}$ is denoted $\mathbf{R}^n$. A function whose domain is a point is essentially a single real number, so the set $\mathbf{R}^1$ of functions $\{1\} \to \mathbf{R}$ may be viewed as the ordinary number line. A function $\mathbf{x} : \{1, 2\} \to \mathbf{R}$ is an ordered pair (x_1, x_2), and the set $\mathbf{R}^2$ of all such functions may be viewed as the Cartesian plane; the point $(-2, 3)$ is the function defined by $\mathbf{x}(1) = -2$ and $\mathbf{x}(2) = 3$, for example. Similarly, the set of real-valued functions on $\underline{\mathbf{3}} = \{1, 2, 3\}$ may be regarded as Cartesian space. There is no mathematical reason to stop at domains with three points, but the spaces of functions become difficult to visualize.

The remarks above hide a subtle point. If $A \subset \mathbf{R}$ is an infinite set, say the closed unit interval $[0, 1]$, then the set of real-valued functions on A is absolutely vast; roughly, there is one coordinate axis for each element of A, and these are in some sense mutually perpendicular! On the other hand, a *single element* of this set (that is, a function $f : A \to \mathbf{R}$) can be pictured as a graph in the plane. In other words, *the set of graphs in the plane is vast.* One says that a real-valued function on $\underline{\mathbf{n}}$ depends on "finitely many parameters" or that the space of real-valued functions on $\underline{\mathbf{n}}$ "has n degrees of freedom." Stretching language a bit, a single function $f : [0, 1] \to \mathbf{R}$ depends on infinitely many parameters, and the space of real-valued functions on $[0, 1]$ has infinitely many degrees of freedom.

Sequences

Let X be a non-empty set. A function $a : \mathbf{N} \to X$ is called a *sequence in X*, and is also denoted $(a_k)_{k=0}^{\infty} \subset X$. Conceptually, a sequence is an

infinite ordered list of (possibly non-distinct) points in X. As suggested by the notation, an ordered n-tuple is just a finite sequence.

A sequence of numbers may be defined by a formula, such as $a_k = 1/(k+1)$ or $a_k = (-1)^k$, or by a recursive specification as in

$$a_0 = 2, \qquad a_{k+1} = \frac{1}{2}\left(a_k + \frac{2}{a_k}\right) \quad \text{for } k \geq 0. \tag{3.9}$$

(Compare Lemma 2.29; this sequence gives successively better approximations to $\sqrt{2}$.) In practice, sequences often arise recursively, and finding a closed formula is highly desirable (if sometimes difficult or impossible). Further examples are given in the exercises.

Sequences are among the most important technical tools of calculus. Cantor's construction of the real field is based on sequences of rational numbers. Generally, if one wants to study an object, perhaps an irrational number like π, a function like cos, or the area enclosed by a curve in the plane, a natural approach is to consider a sequence that "approximates" the target object in some sense. The hope is to use properties of the approximators to deduce properties of the "limit" object.

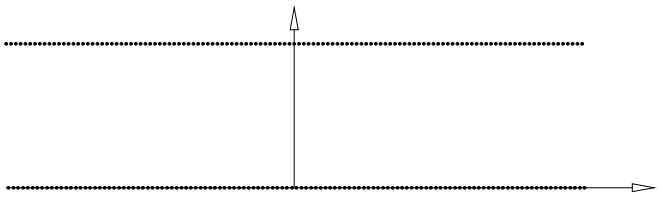
"Pathological" Examples

Very strange functions can be specified by collections of rules or formulas. The important thing is that to *every* point of the domain must be associated *exactly one* value.

Example 3.10 The characteristic function of $\mathbf{Q}$ in $\mathbf{R}$ is defined by

$$\chi_{\mathbf{Q}}(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$$

Because the rationals are dense in the reals, see Theorem 2.33, the graph looks something like



Descriptively, the graph is like two horizontal lines, with the understanding that each vertical line only hits the graph at one point. Of course, the actual graph contains infinitely fine detail, and unlike in the picture there are no "consecutive" points. □

Example 3.11 Every real number is either rational or irrational, and every rational number has a *unique* representation as p/q in lowest terms. Define a function $f : \mathbf{R} \to \mathbf{R}$ by

$$f(x) = \begin{cases} 1/q & \text{if } x = p/q \text{ in lowest terms,} \\ 0 & \text{if } x \text{ is irrational.} \end{cases} \tag{3.10}$$

Figure 3.13 depicts the graph of f; a fully accurate picture is impossible because printed points cannot resolve the arbitrarily fine details in the graph. The white band near the horizontal axis results because no points with $q > 40$ are shown. □

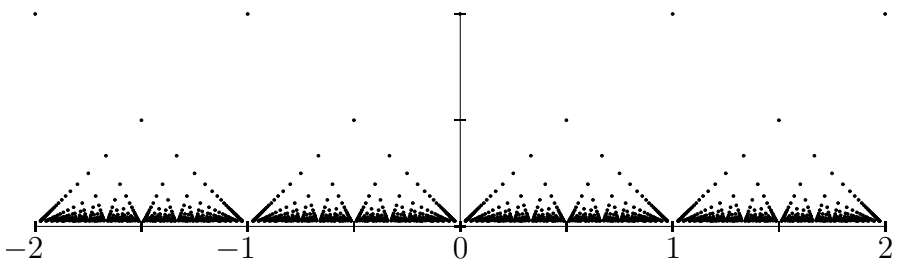


Figure 3.13: The denominator function.

Example 3.12 By Exercise 2.18, every real number has a decimal expansion, and this expansion is unique if we agree to write non-zero terminating decimals (such as 0.25) with an infinite string of 9's instead (as in $0.249\overline{9}$). With this convention in mind (and writing "x" for "the decimal expansion of x"), define $f : (0, 1) \to \mathbf{Z}$ by

$$f(x) = \begin{cases} k & \text{if the digit 7 occurs exactly } k \text{ times in } x, \\ -1 & \text{if the digit 7 occurs infinitely many times in } x. \end{cases}$$

Since every real x has a unique decimal expansion, the number of occurrences of the digit 7 is well-defined. However, for a specific choice of x, it is likely to be impossible to calculate $f(x)$; the value $f(\pi - 3)$ is believed to be -1, but this is not known.

To convey how "chaotically" the values of f are distributed, we sketch an argument that in *every* open interval (a, b)—no matter how small—the function f achieves arbitrarily large values. In the interval $(0, 10^{-10})$, for instance, we find the numbers with decimal expansion $x = 0.0\cdots07\cdots79\ldots$ having (at least) ten zeroes, followed by a string of k 7's and an infinite string of 9's; for this number, $f(x) = k$. It should be clear that essentially the same reasoning applies to an arbitrary interval. The graph of this function can be represented as a collection of horizontal lines, at height -1, 0, 1, 2, and so forth, subject to the proviso that the lines are not really "solid": Each vertical line intersects the graph exactly once. □

3.3 Composition, Iteration, and Inverses

If $f : X \to Y$ and $g : Y \to Z$ are functions (particularly, the image of f is contained in the domain of g), then the *composition of g and f* is the function $g \circ f : X \to Z$, read "g of f," defined by $(g \circ f)(x) = g\big(f(x)\big)$ for all $x \in X$. In words, apply f to x, then apply g to the output. Composition of functions is associative, but is not generally commutative; one composition may be defined while the other is not, but even if both compositions are defined, they are usually unrelated. For example, if $f(x) = x+1$ ("adding one") and $g(x) = x^2$ ("squaring"), then

$$(g \circ f)(x) = g(x+1) = (x+1)^2 = x^2 + 2x + 1,$$
$$(f \circ g)(x) = f(x^2) = x^2 + 1.$$

Iteration

If $f : X \to X$, that is, the domain and range of f are the same set, then f may be composed with itself, over and over. (It is not necessary that f be onto; why not?) The nth *iterate* of f is the function $f^{[n]}$ defined by "composing f with itself n times." The formal recursive definition is

$$f^{[0]} = I_X, \qquad f^{[n+1]} = f \circ f^{[n]} \quad \text{for } n \geq 0. \tag{3.11}$$

Thus $f^{[1]} = f$, $f^{[2]} = f \circ f$, $f^{[3]} = f \circ f \circ f$, and so on.

The sequence defined by equation (3.9) is obtained by iterating

$$f(x) = \frac{1}{2}\left(x + \frac{2}{x}\right)$$

regarded as a mapping from $(0, +\infty)$ to itself. In this example, $x_0 = 2$, and $x_{k+1} = f(x_k)$ for $k \geq 0$. Generally, let $f : X \to X$, and let an "initial point" $x_0 \in X$ be given. The sequence $(x_n)_{n=1}^{\infty}$ defined by $x_k = f^{[k]}(x_0)$ consists of subsequent locations of x_0 under repeated application of the mapping f (its "forward history"), and the set $\{x_k\}_{k=0}^{\infty} \subset X$ is the *orbit of* x_0 *under iteration of* f. Simultaneous consideration of all points of X gives a *discrete dynamical system*; the set X is regarded as a "space" and its points are "mixed" by the mapping f. Of special interest are points $x \in X$ with $f(x) = x$, the *fixed points* of f. We will see why in Chapter 4.

Inversion

If a function $f : X \to Y$ is regarded as an operation or procedure, it is often desirable to "undo" the effect of f by applying another function, that is, to recover the input x for each output value $y = f(x)$. Not every function is amenable to "inversion". A function $f : X \to Y$ is said to be *invertible* if there exists a function $g : Y \to X$ such that

$$f \circ g = I_Y, \quad \text{that is, } (f \circ g)(y) = y \text{ for all } y \in Y, \tag{3.12}$$

and

$$g \circ f = I_X, \quad \text{that is, } (g \circ f)(x) = x \text{ for all } x \in X. \tag{3.13}$$

These two equations are "dual" to each other in the sense that simultaneously exchanging f with g and X with Y converts each equation into the other. Also, they are logically independent, and the properties they specify have already been encountered:

Proposition 3.13. *Let $f : X \to Y$ be a function. There exists a function g satisfying equation (3.12) if and only if f is onto, and there exists a function satisfying equation (3.13) if and only if f is one-to-one.*

The proof amounts to reformulation of the definitions; you will learn more by trying to prove this yourself than you will simply by reading the proof below. To understand intuitively what the proposition means, consider an analogy. Suppose you are moving to a new apartment, and have bought printed labels ("kitchen", "bathroom", "garage", etc.) from the moving company. You have several types of possessions (dishes, glasses, towels, clothes, books...), and each item gets put into a box that bears a label. Your mathematician friend is packing boxes while

you run errands: Each box is labeled according to the room in which its contents belong.

The types of item to be packed are points of X, the types of labels are points of Y, and your friend's labeling scheme is a function f. Your goal is to identify your possessions by looking only at the labels on the boxes, namely to recover points of X from their function values. In this situation, the two halves of Proposition 3.13 address the following questions:

- For each type of room label, is there a corresponding box? A function g as in (3.12) exists iff the labeling function is surjective.
- Can you determine each box's contents just by looking at the label? A function g as in (3.13) exists iff the labelling function is injective.

Proof. If there is a function satisfying equation (3.12), then every $y \in Y$ is in the image of f, since the element $x := g(y) \in X$ is mapped to y by f. Conversely, if f is onto, then for each $y \in Y$, the set $f^{[-1]}(\{y\})$ is non-empty (by definition). So for each $y \in Y$, it is possible to pick an element $x \in f^{[-1]}(\{y\})$, and a family of such choices is nothing but a function $g : Y \to X$. By construction, equation (3.12) holds.

Now suppose f is one-to-one. Pick an element $x_0 \in X$ arbitrarily, and define the function $g : Y \to X$ by

$$g(y) = \begin{cases} x & \text{where } f(x) = y, \text{ if } y \in f(X), \\ x_0 & \text{otherwise.} \end{cases}$$

This prescription defines a function (i.e., is single-valued) because f is one-to-one, and it is clear that for this function g, equation (3.13) holds. Conversely, suppose equation (3.13) holds, and let x_1 and x_2 be elements of X for which $f(x_1) = f(x_2)$. We want to show that $x_1 = x_2$. But this is clear, since

$$x_1 = g\big(f(x_1)\big) = g\big(f(x_2)\big) = x_2$$

by equation (3.12). □

A function g satisfying equation (3.12) is called a *right inverse* of f, or a *branch of* f^{-1}, while a function g satisfying equation (3.13) is a *left inverse* of f. Left inverses arise rarely, because you can replace the range of a function with its image, whereupon the function becomes

surjective. By contrast, branches of f^{-1} arise naturally in algebra and trigonometry.

A single function may have several left inverses or several right inverses; however, if a function f has *both* a left inverse g and a right inverse h, then $g = h$:

$$g = g \circ (f \circ h) = (g \circ f) \circ h = h.$$

In this event, Proposition 3.13 shows that f is a bijection. Thus, a bijection and an invertible function are the same thing. Conceptually, a bijection is nothing but a renaming: elements of X are objects of interest, while elements of Y are "labels," and a bijection $f : X \to Y$ associates a unique label to each object, and a unique object to each label. Bijections between infinite sets can look strange at first. The following examples give a small sampling of interesting bijections.

Example 3.14 The identity map $I_X : X \to X$ is invertible for every non-empty set X, and is its own inverse. If $a \in \mathbf{R}$, then the function $x \mapsto x + a$, called *translation by* a, is a bijection, whose inverse is translation by $-a$. If $a \neq 0$, then the function $x \mapsto ax$, called *scaling by* a, is a bijection whose inverse is scaling by $1/a$. The analogous functions are bijections in an arbitrary field. Every one-to-one function is a bijection onto its image. For example, the function $f : \mathbf{Z} \to \mathbf{Z}$ defined by $f(n) = 2n$ is a bijection between the set of integers and the set of even integers. Observe that an infinite set can be put in one-to-one correspondence with a *proper* subset of itself. □

Example 3.15 A *logarithm* is a bijection $L : (0, \infty) \to \mathbf{R}$ such that

$$L(xy) = L(x) + L(y) \qquad \text{for all } x, y \in \mathbf{R}.$$

The existence of logarithms is deduced in Chapter 12. Historically, logarithms were important because they convert multiplication into addition, provided there is an effective means of going between $x \in (0, \infty)$ and $L(x) \in \mathbf{R}$. Before the age of electronic computers, the conversion was done by means of logarithm tables and slide rules. Logarithms are of great importance in pure mathematics, the sciences, and engineering; stellar magnitudes, loudness (measured in decibels), and acidity (pH) are all measured using logarithmic scales. □

Example 3.16 There are calculational methods for finding inverses of functions defined by formulas. In high school courses the usual prescription is to "exchange x and y in the equation $y = f(x)$, and then

solve for y." Equivalently, solve $y = f(x)$ for x. This is essentially correct, though care must be taken with domains and ranges, as this example illustrates.

Let $f : [-1, 0] \to [0, 1]$ be defined by $f(x) = x^2$. This function is one-to-one and onto. Formal solution for x gives $x = \pm\sqrt{y}$. This "equation" (really a pair of equations) does not determine f^{-1}, though it narrows down the possibilities enough that the inverse can be found by inspection. Because the domain of f is $[-1, 0]$, the range of f^{-1} must be this same interval. Therefore, $f^{-1} = -\sqrt{\ }$, since by definition the square root function takes non-negative values. □

Example 3.17 The "obvious" bijection between the set $\{a, \ldots, z\}$ and the set $\{1, \ldots, 26\} \subset \mathbf{N}$ can be used to encode and transmit messages as numbers. Decoding a message amounts to inverting the bijection that encoded the message originally. A more sophisticated code would allow for both capital and lowercase letters, punctuation, and numerals. The so-called *ASCII* character encoding (known as "ISO 8859-I" outside the United States) is just such a correspondence, and is widely used for text storage. □

Inversion of Monotone Functions

A strictly monotone function is injective, hence is a bijection to its image. If f is increasing, then f^{-1} is also increasing: Let $y_1 < y_2$ be elements of the image, and let $x_i = f^{-1}(y_i)$. One of the inequalities $x_1 < x_2$ or $x_1 > x_2$ must hold. Because f is increasing, the second possibility cannot occur. Thus, if $y_1 < y_2$, then $f^{-1}(y_1) < f^{-1}(y_2)$. The same argument proves that if f is decreasing, then f^{-1} is decreasing. Note well that an injective function is generally not monotone.

Permutations and Cardinality

Recall that for $n \in \mathbf{N}$, the corresponding initial segment $\underline{\mathbf{n}}$ is the set $\{1, \ldots, n\}$. A bijection from $\underline{\mathbf{n}}$ to itself is called a *permutation on n letters*. There are $n!$ permutations on n letters. It is fairly clear intuitively (and can be proven by mathematical induction) that there exists an injection $i : \underline{\mathbf{n}} \to \underline{\mathbf{m}}$ if and only if $n \leq m$.

G. Cantor's idea for comparing the "sizes" of infinite sets generalizes this; two sets X and Y have the same *cardinality* if there is a bijection $f : X \to Y$. More generally, "the cardinality of X is no larger than

the cardinality of Y" iff there is an injection $i : X \to Y$. (By Proposition 3.13, this is equivalent to existence of a surjection $p : Y \to X$.) As in Example 3.14, the cardinality of an infinite set can be the same as that of a proper subset. By definition, a set X is *countable* if there exists a bijection $f : \mathbf{N} \to X$, and is *at most countable* if either finite or countable. Cantor believed at first that all infinite sets are countable. Later he proved the contrary, both by a general argument (see Theorem 3.18) and in a spectacular special case (Theorem 3.19). Cantor's work met with acrimonious disapproval from several mathematicians of the late 19th Century, but is now known to be fundamentally sound.

Theorem 3.18. *Let X be a set, and let $\mathcal{P}(X)$ be its power set, the set of all subsets of X. Then there is no surjection $p : X \to \mathcal{P}(X)$; every set has strictly smaller cardinality than its power set.*

Proof. Cantor showed that if $p : X \to \mathcal{P}(X)$ is an arbitrary function, then there exists an element of $\mathcal{P}(X)$ that is not in the image. The mapping p associates to each $x \in X$ a *subset* $p(x) \subset X$. For each x, it makes sense to ask whether or not $x \in p(x)$, and the answer is either "yes" or "no" (depending on x and p). Consider the set

$$A = \{x \in X \mid x \notin p(x)\} \subset X;$$

the set $A \in \mathcal{P}(X)$ depends on p, but is unambiguously defined.

For each $x \in X$, either $x \in A$, or $x \notin A$. If $x \in A$, then $x \notin p(x)$, so $p(x) \neq A$ as sets (one contains x and the other doesn't). On the other hand, if $x \notin A$, then $x \in p(x)$, and again $p(x) \neq A$. In summary, if $x \in X$, then $p(x) \neq A$, namely, p is not surjective. □

This theorem shows that for every set—possibly infinite—there is another set with strictly larger cardinality! One can perhaps sympathize with those mathematicians who felt that only madness or linguistic fog (infinite progressions of larger and larger infinities) lay in this direction. The following theorem, again due to Cantor, shows that there are "more" irrational numbers than rational numbers.

Theorem 3.19. *The set of rational numbers is countable; the set of real numbers is not countable. Specifically, there is a bijection $f : \mathbf{N} \to \mathbf{Q}$, but there does not exist a surjection $p : \mathbf{N} \to \mathbf{R}$.*

Proof. (Sketch) Conceptually, a bijection $f : \mathbf{N} \to \mathbf{Q}$ is a method of listing all the elements of $\mathbf{Q}$. We first construct a surjection from $\mathbf{N}$

to the set of pairs (p, q) of integers with $q > 0$, see Figure 3.14, then "strike off" pairs that are not in lowest terms. This gives the desired bijection. Note carefully that the ordering of $\mathbf{Q}$ by $<$ does *not* give a bijection, since there is no such thing as a pair of "consecutive" rational numbers.

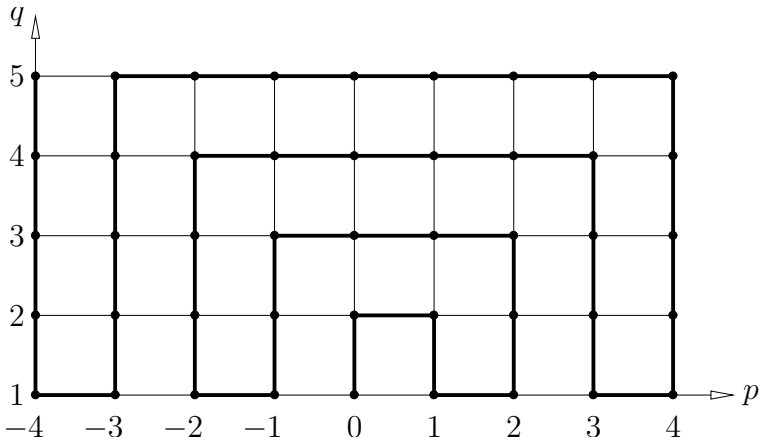


Figure 3.14: Constructing a bijection from $\mathbf{N}$ to $\mathbf{Q}$.

According to most peoples' intuition, there are more rational numbers than natural numbers, because there are infinitely many rationals between each pair of natural numbers. The bijection depicted in Figure 3.14 shows that this intuition is incorrect; when counting infinite sets, the order in which elements are enumerated matters, because an infinite set can be put into bijective correspondence with a proper subset.

To prove that $\mathbf{R}$ is not countable may seem impossible after the argument above; if we fail to find a bijection, perhaps we were simply not clever enough! As you will notice, the argument we use here is completely different. It is enough to show that some subset of $\mathbf{R}$ is not countable. Consider the set X of real numbers that are represented by a decimal containing only the digits 0 and 1, and let $f : \mathbf{N} \to X$ be an arbitrary map. List the elements in the image as in (3.14), which depicts a "typical" choice of f:

$$\begin{aligned} f(1) &= 0.\mathbf{1}101110\ldots \\ f(2) &= 0.0\mathbf{1}00100\ldots \\ f(3) &= 0.11\mathbf{0}1100\ldots \\ f(4) &= 0.101\mathbf{0}010\ldots \end{aligned} \qquad \rightsquigarrow \quad x = .0011\ldots \tag{3.14}$$

To show that f is not onto, it is enough to construct a number x that is not in the image of f. Consider the kth decimal of the kth number $f(k)$; if this decimal is 0 then take the kth decimal of x to be 1 and *vice versa*. Then $x \in X$ since its decimal expansion contains only 0s and 1s, but x is not in the image of f because x and $f(k)$ differ in the kth decimal. We have shown that f is not onto; since f was arbitrary, there is no surjection $f : \mathbf{N} \to X$, *a fortiori* no surjection $\mathbf{N} \to \mathbf{R}$. □

Most people who follow this proof for the first time immediately ask, "Why not add the new number to the list?" To understand why this is an irrelevant point, you must interpret the claim correctly: An *arbitrary* map $f : \mathbf{N} \to X$ is not onto. The function f is specified *before* the number x is constructed. You may well appreciate the feelings of Cantor's detractors, but this theorem is perfectly consistent with the definitions.

Isomorphism

Two mathematical structures that are "abstractly equivalent" are usually regarded as being "the same." For each mathematical structure, there is a concept of *isomorphism* that defines precisely what is meant by "abstract equivalence." Mathematical structures encountered previously include sets, commutative groups, fields, and ordered fields. The next example explains isomorphisms in detail for sets and commutative groups.

Two sets X and Y are *isomorphic* if and only if there exists a bijection $\phi : X \to Y$. Intuitively, a set has no attributes other than "the number of elements" (which may be finite or infinite). The map ϕ is an *isomorphism* between X and Y, and as mentioned above "renames" elements of X. The sets

$$X = \{0, 1\}, \qquad Y = \big\{\varnothing, \{\varnothing\}\big\}, \qquad \text{and} \quad Z = \{\text{True}, \text{False}\}$$

are mutually isomorphic. If X and Y are isomorphic sets having more than one element, then there are many isomorphisms between them, and there is usually no reason to select one over another. If $X = Y$, however, then the identity map I_X is in a sense the "natural" choice of isomorphism. You might say[3] that two sets with the same number of elements are isomorphic, but some sets are more isomorphic than others.

[3]With apologies to G. Orwell.

The situation is similar, but more interesting, when considering sets with additional structure. In this case, an "isomorphism" should "preserve the additional structure." Suppose $(\mathbf{G}, +)$ and $(\mathbf{H}, \oplus)$ are commutative groups. This means that $\mathbf{G}$ is a non-empty set, and that $+$ is a way of "adding" two elements of $\mathbf{G}$ to get a third, subject to axioms A.1–A.4 on page 51. Similar comments hold for $\mathbf{H}$ and $\oplus$. An *isomorphism* between $(\mathbf{G}, +)$ and $(\mathbf{H}, \oplus)$ is a bijection $\phi : \mathbf{G} \to \mathbf{H}$ such that

$$\phi(x + y) = \phi(x) \oplus \phi(y) \qquad \text{for all } x,\, y \in \mathbf{G}. \tag{3.15}$$

Equation (3.15) says that the operations $+$ and $\oplus$ correspond under ϕ; adding in $\mathbf{G}$ and then relabelling (the left-hand side) is the same as relabelling and then adding in $\mathbf{H}$ (the right-hand side). As far as properties of commutative groups are concerned, $(\mathbf{G}, +)$ and $(\mathbf{H}, \oplus)$ are indistinguishable. Their sets of elements and/or their operations of "addition" may look very different, but abstractly the structures are the same. A logarithm (Example 3.15) is an isomorphism L between the multiplicative group of positive real numbers, $(\mathbf{R}^+, \cdot)$, and the additive group of real numbers, $(\mathbf{R}, +)$.

The concept of isomorphism extends to more complicated mathematical structures in a straightforward way. An ordered field $(\mathbf{F}, +, \cdot, <)$ consists of a non-empty set $\mathbf{F}$, two operations $+$ and $\cdot$, and a relation $<$ on $\mathbf{F}$ subject to axioms. The ordered field $(\mathcal{F}, \oplus, \odot, \prec)$ is *isomorphic* to $(\mathbf{F}, +, \cdot, <)$ (as an ordered field) if there exists a bijection $\phi : \mathbf{F} \to \mathcal{F}$ such that analogues of equation (3.15) hold for the arithmetic operations, and such that the order relations correspond in the sense that $x < y$ if and only if $\phi(x) \prec \phi(y)$. As above, isomorphic ordered fields are abstractly indistinguishable, so far as questions about ordered fields are concerned.

Once these concepts are understood, it is possible to make the statement of Theorem 2.30 ("existence and uniqueness of the real numbers") precise. First of all, there exists a complete, ordered field $(\mathbf{R}, +, \cdot, <)$. To say "$\mathbf{R}$ contains $\mathbf{Q}$" means there is an injection $i : \mathbf{Q} \to \mathbf{R}$ that is an isomorphism (of ordered fields) onto its image. Uniqueness means that every *complete* ordered field $(\mathcal{R}, \oplus, \odot, \prec)$ is isomorphic to $(\mathbf{R}, +, \cdot, <)$ as an ordered field.

3.4 Linear Operators

A great deal of conceptual economy is obtained by introducing some terminology from linear algebra. The fundamental operations of calculus—integration and differentiation—may be treated as "functions" whose domains are *spaces of functions*.

Vector Spaces

Let $X \subset \mathbf{R}$ and let $\mathscr{F}(X, \mathbf{R})$ denote the set of real-valued functions on X. When X is a finite set, the space $\mathscr{F}(X, \mathbf{R})$ of functions is (essentially) $\mathbf{R}^n$ and we regard the general element as "a list of real numbers indexed by points of X".

The operations of interest to us are addition of functions and "scalar multiplication". If f and g are elements of $\mathscr{F}(X, \mathbf{R})$ and if c is a real number, then we define functions $f + g$ and $cf \in \mathscr{F}(X, \mathbf{R})$ by

$$\begin{aligned} (f+g)(x) &= f(x) + g(x) \\ (cf)(x) &= c \cdot f(x) \end{aligned} \qquad \text{for all } x \in X. \tag{3.16}$$

The set $\mathscr{F}(X, \mathbf{R})$ together with these algebraic operations is an example of a *vector space*. There is an axiomatic definition similar to the definition of a field, which you will encounter in a linear algebra course.

A non-empty subset $V \subset \mathscr{F}(X, \mathbf{R})$ is a *vector subspace* if two conditions hold:

- (Closure under addition) If f and $g \in V$, then $f + g \in V$
- (Closure under scalar multiplication) If $f \in V$, then $cf \in V$ for all $c \in \mathbf{R}$

For example, the set of polynomial functions on X is a vector subspace of $\mathscr{F}(X, \mathbf{R})$, as is the set of step functions. The set of indicator functions on X is *not* a vector subspace: The sum of two indicator functions is not generally an indicator.

Linear Mappings

Let V and W be vector subspaces of $\mathscr{F}(X, \mathbf{R})$. A mapping $L : V \to W$ takes a function f as input and returns a function Lf as output. It is

customary to write Lf instead of $L(f)$ to avoid excessive parentheses. A mapping $L : V \to V$ is called an *operator* on V.

A mapping $L : V \to W$ is *linear* if

$$\begin{aligned} L(f+g) &= Lf + Lg \\ L(cf) &= c \cdot Lf \end{aligned} \qquad \text{for all } f \text{ and } g \text{ in } V, \text{ all real } c. \tag{3.17}$$

You may regard a linear mapping as one that "respects the vector space structure". A *linear functional*[4] is a linear mapping $T : V \to \mathbf{R}$.

Example 3.20 Fix $a \in \mathbf{R}$ and define an operator $L_a : \mathscr{F}(\mathbf{R}, \mathbf{R}) \to \mathscr{F}(\mathbf{R}, \mathbf{R})$ by $L_a f(x) = f(x - a)$. The effect of L_a is to "shift" f to the right by a in the domain. Linearity is immediate, as you should check. □

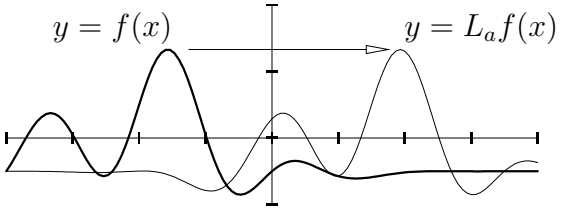


Figure 3.15: The translation operator.

Example 3.21 A closely related example is the *reflection operator* R, defined by $Rf(x) = f(-x)$. Geometrically, R reflects the graph of f about the vertical axis. □

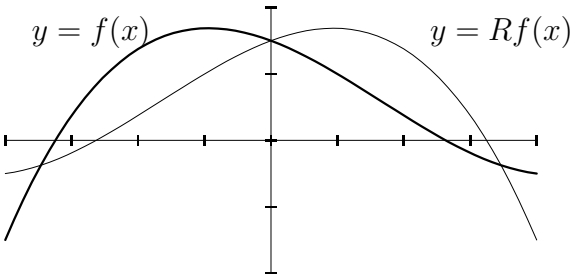


Figure 3.16: The domain-reflection operator.

[4] In mathematics, "functional" is a noun!

Example 3.22 Fix $x \in X$; the *evaluation functional* $\mathrm{ev}_x : \mathscr{F}(X, \mathbf{R}) \to \mathbf{R}$ is defined by $\mathrm{ev}_x(f) = f(x)$, namely "evaluation of f at x". The definition of addition and scalar multiplication of functions says that ev_x is a linear functional. □

The operator S defined by $Sf(x) = f(x)^2$ is not linear; for example, multiplying f by 2 and applying S multiplies the output by $2^2 = 4$.

Symmetries of Functions

The reflection and translation operators introduced above lead us to some interesting classes of functions.

Even and Odd Functions

Let $A \subset \mathbf{R}$ be an interval of the form $[-a, a]$ for some $a > 0$. A function $f : A \to \mathbf{R}$ is *even* if

$$f(-x) = f(x) \qquad \text{for all } x \in A, \tag{3.18}$$

and is *odd* if

$$f(-x) = -f(x) \qquad \text{for all } x \in A, \tag{3.19}$$

see Figure 3.17. The terminology arises from monomial functions $x \mapsto$

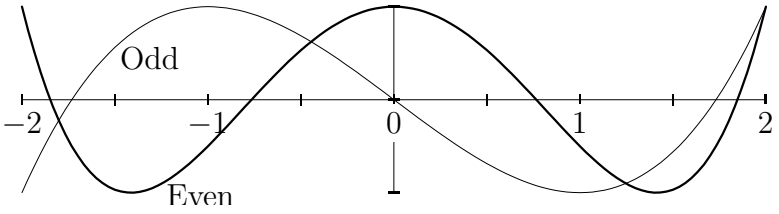


Figure 3.17: Even and odd functions.

x^k for k a positive integer; the "parity" of a monomial in the above sense is the same as the parity of the exponent k as an integer, since

$$(-x)^k = (-1)^k x^k = \begin{cases} x^k & \text{if } k \text{ is even,} \\ -x^k & \text{if } k \text{ is odd.} \end{cases}$$

Evenness and oddness are "global" properties: They depend on the behavior of f on the entire domain.

There is a beautiful interpretation in terms of the reflection operator of Example 3.21: A function is even iff $Rf = f$ (f is *invariant* under R) and is odd iff $Rf = -f$ (f is *anti-invariant* under R). For every function f, the operator R exchanges f and Rf, since $R(Rf) = f$.

Lemma 3.23. *The spaces of even and odd functions on $A = [-a, a]$ are vector subspaces of $\mathscr{F}(A, \mathbf{R})$.*

Proof. This is a restatement of linearity of R: If f and g are even and c is real, then

$$R(f + g) = Rf + Rg = f + g, \qquad R(cf) = c \cdot Rf.$$

Thus $f + g$ and cf are even, so the set of even functions is a vector subspace of $\mathscr{F}(A, \mathbf{R})$. The proof for odd functions is essentially identical. □

Since a sum of even functions is even, a polynomial is even if *every term* has even degree. The converse is also true, see Proposition 3.24 below. These remarks are true if "even" is replaced everywhere by "odd."

Every constant function on $\mathbf{R}$ is even. The only constant function that is odd is the zero function; in fact, the zero function is easily seen to be the only function that is both even and odd. The "signum" function

$$\operatorname{sgn}(x) = \begin{cases} \dfrac{x}{|x|} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases} = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{3.20}$$

is odd.

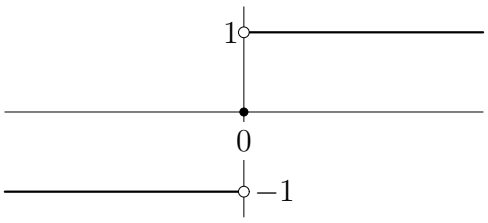


Figure 3.18: The signum function.

Most functions are neither even nor odd. However, every function f on a symmetric domain can be expressed (uniquely) as the sum of an

even function f_{even} and an odd function f_{odd}. Indeed, the functions defined by

$$\begin{aligned} f_{\text{even}}(x) &= \tfrac{1}{2}\big(f(x) + f(-x)\big) \\ f_{\text{odd}}(x) &= \tfrac{1}{2}\big(f(x) - f(-x)\big) \end{aligned} \tag{3.21}$$

are easily shown to have the required properties. These formulas are arrived at by writing $f = f_{\text{even}} + f_{\text{odd}}$ and using equations (3.18) and (3.19). Observe that f is even exactly when its odd part f_{odd} is the zero function, and that f is odd iff its even part is identically zero. In terms of R, equation (3.21) says

$$f_{\text{even}} = \tfrac{1}{2}(f + Rf), \qquad f_{\text{odd}} = \tfrac{1}{2}(f - Rf).$$

To obtain an even function from f we average f and Rf, and to obtain an odd function we average f and $-Rf$: The even and odd parts of f are obtained by "weighted averaging over the action of R".

To complete the discussion of parity of functions, we characterize even and odd polynomials.

Proposition 3.24. *Let $p : \mathbf{R} \to \mathbf{R}$ be a polynomial function. Then p is even iff every term of p has even degree, and p is odd iff every term has odd degree.*

Concisely (if less transparently), p is even iff there exists a polynomial q with $p(x) = q(x^2)$ for all $x \in \mathbf{R}$, and p is odd iff there exists a polynomial q with $p(x) = x\,q(x^2)$ for all $x \in \mathbf{R}$.

Proof. Suppose p is a polynomial, and let p_{e} and p_{o} denote the sum of the even-degree terms and the sum of the odd-degree terms. As noted previously these polynomial functions are respectively even and odd, and their sum is p. They must be the even and odd parts of p by uniqueness. □

Periodic Functions

Let ℓ be a non-zero real number. A function $f : \mathbf{R} \to \mathbf{R}$ is *periodic with period ℓ*—or *ℓ-periodic*—if

$$f(x - \ell) = f(x) \qquad \text{for all } x \in \mathbf{R}. \tag{3.22}$$

In terms of the translation operator T_ℓ from Example 3.20, f is ℓ-periodic iff $T_\ell f = f$.

By induction, $f(x+n\ell) = f(x)$ for all $n \in \mathbf{Z}$; consequently the graph of an ℓ-periodic function consists of "waveforms" of length ℓ, repeated *ad infinitum*. The restriction of an ℓ-periodic function to an interval of length ℓ is called a *period*. Clearly a periodic function is completely specified by each of its periods. Conversely, given a function on a half-open interval of length ℓ, there is a unique *periodic extension* to an ℓ-periodic function.

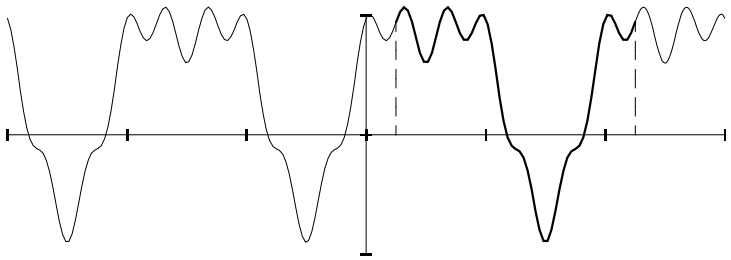


Figure 3.19: A periodic function, and one complete period.

Example 3.25 The *Charlie Brown* function $\mathrm{cb} : \mathbf{R} \to \mathbf{R}$ is the periodic extension of the absolute value function on $[-1, 1)$, see Figure 3.20. Note that cb is piecewise polynomial, in fact, piecewise linear. □

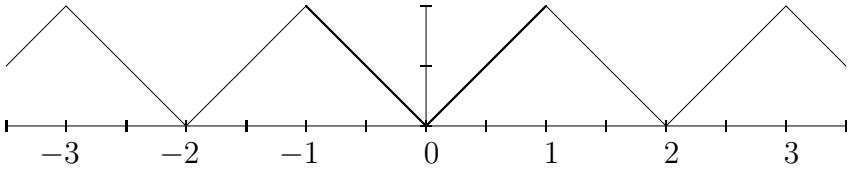


Figure 3.20: The Charlie Brown function.

Positive and Negative Parts

Let f be a real-valued function whose domain is an arbitrary set X. The *positive part* of f is the function $f_+ : X \to \mathbf{R}$ defined by

$$f_+(x) = \max\big(f(x), 0\big) = \begin{cases} f(x) & \text{if } f(x) \geq 0 \\ 0 & \text{if } f(x) < 0 \end{cases}$$

Similarly, the *negative part* of f is defined by $f_-(x) = \min\big(f(x), 0\big)$. You should sketch the positive and negative parts of the function in Figure 3.19 to ensure you understand the definition.

Note that $f(x) = f_+(x) + f_-(x)$ for all x; every real-valued function is a difference of non-negative functions. This observation has amusing and important applications later. For example, we will be able to show that functions in a certain large class can be written as a difference of monotone functions.

Exercises

Exercise 3.1 After typing a long letter, you realize that you have systematically exchanged the words "there" and "their." Luckily your text editor can replace all occurrences of a string with another string. You first replace "there" with "their," and then replace "their" with "there." Does this have the desired effect? Interpret the consequences of these replacements as functions from the set {there, their} to itself. Are these functions one-to-one? How could you successfully exchange all occurrences of "there" and "their" using replacement? ◇

Exercise 3.2 Prove Proposition 3.3. You must establish three inclusions of sets, using only the definitions of functions and preimages. ◇

Exercise 3.3 Give an example of a function $f : X \to Y$ and a specific $A \subset X$ such that the inclusion $A \subset f^{[-1]}\big(f(A)\big)$ is proper.
Hints: Your function must not be one-to-one. Figure 3.8 may help. ◇

Exercise 3.4 Let A and B be subsets of $\mathbf{R}$, and let χ_A and χ_B be their indicator functions, equation (3.7). Establish the following:

(a) $1 - \chi_A = \chi_{(\mathbf{R}\setminus A)}$, the indicator of A^c.

(b) $\min(\chi_A, \chi_B) = \chi_{A\cap B} = \chi_A \cdot \chi_B$.

(c) $\max(\chi_A, \chi_B) = \chi_{A\cup B} = \chi_A + \chi_B - \min(\chi_A, \chi_B)$.

(d) $\chi_A + \chi_B \pmod 2 = \chi_{A\triangle B}$. (See Exercise 1.2)

In words, Boolean operations on sets and characteristic functions are closely related. ◇

Exercise 3.5 This exercise characterizes step functions.

(a) For $k = 1, \ldots, n$, let I_k be an interval, χ_k the characteristic function of I_k, and c_k a real number. Use Exercise 3.4 and induction on n to prove that

$$(\ddagger) \qquad f(x) = \sum_{k=1}^{n} c_k \chi_k(x)$$

is a step function on $\mathbf{R}$. The difference between this and equation (3.8) is that the intervals need not be pairwise disjoint here.

(b) Can every step function be written in the form (‡)? If so, prove it; if not, what properties of the sets I_k need to be modified?

(c) Is the representation from part (b) unique? If so prove it; if not, what properties of the sets I_k need to be modified?

It may help to sketch some functions of the form $c_1\chi_1 + c_2\chi_2$ for which the sets I_1 and I_2 are, or are not, disjoint. ◇

Exercise 3.6 In each part, $f : X \to Y$ is a function and $A \subset X$. Determine whether each of the following implications is valid (give a proof) or not (find a counterexample).

(a) If f is injective, then $f|_A$ is injective.

(b) If $f|_A$ is injective, then f is injective.

(a) If f is surjective, then $f|_A$ is surjective.

(a) If $f|_A$ is surjective, then f is surjective.

It may help to consider contrapositives. ◇

Rational and Algebraic Functions

Exercise 3.7 Let $S^1 \subset \mathbf{R}^2$ denote the unit circle, and let $X \subset S^1$ be the complement of the point $(0, 1)$. Define a function $p : \mathbf{R} \to X$ as in Figure 3.21. Geometrically, join $(0, 1) \in S^1$ to $(t, 0)$ by a straight line, and let $p(t) = (x, y)$ be the point of intersection with the circle.

(a) Use similar triangles to find a formula for (x, y) in terms of t.

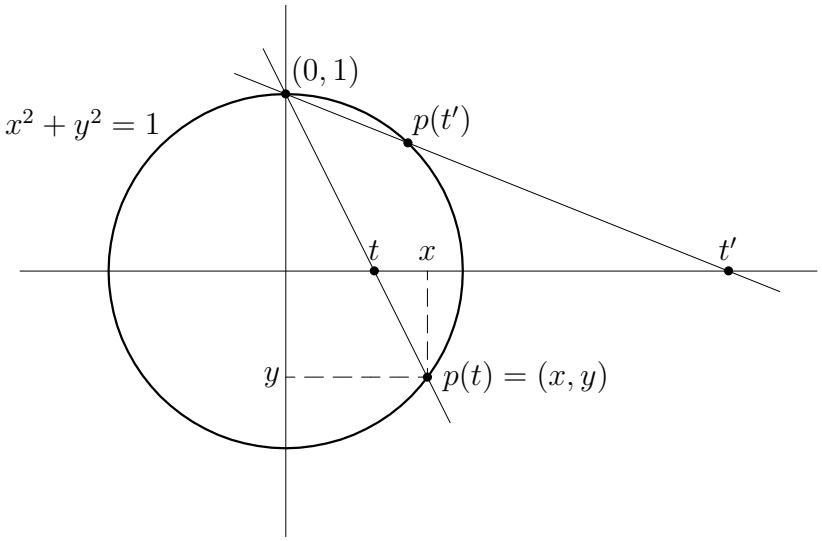


Figure 3.21: Stereographic projection.

(b) Show that p is one-to-one and onto, both geometrically and algebraically. Find a formula for p^{-1} (i.e., express t in terms of x and y.) The mapping p^{-1} is called *stereographic projection.*

(c) Use part (b) to prove that t is rational iff x and y are rational. Thus, stereographic projection characterizes "rational points" on the circle.

(d) Show that under stereographic projection, the mapping $f(t) = 1/t$ corresponds to reflection of the circle through the horizontal axis. If you can, give both an algebraic and a geometric proof.

(e) Show that the rational mapping $f(t) = (t-1)/(t+1)$ corresponds to a one-quarter rotation counterclockwise of the circle.
Hint: The rotation maps $(x, y) \mapsto (-y, x)$.

Part (d) suggests that one might say $1/0 = \infty$ and $1/\infty = 0$. Compare with the section on projective infinite limits in Chapter 4. ◇

Exercise 3.8 Prove that every rational function is algebraic. (Formally this is trivial, but be sure to account for the exact definitions, including domains and ranges.) ◇

Exercise 3.9 Let $F(x, y) = 1+y+xy+xy^2$. Find all algebraic functions implicitly defined by F, and sketch the zero locus $Z(F)$. (Suggestion: Use the quadratic formula.) ◇

Exercise 3.10 Let $F(x, y) = (x^2 + y^2)^2 - (x^2 - y^2)$. Find all algebraic functions defined by F, and locate their graphs in Figure 3.12. ◇

Exercise 3.11 Sketch the loci $x(1 + x)(k + x) - y^2 = 0$ for $k = -1$, 0, and 1. (It may help to sketch the graph $y = x(1 + x)(k + x)$ first.) ◇

Exercise 3.12 Let $\mathbf{F}$ be a field, and let p and d be non-constant polynomials over $\mathbf{F}$, with $\deg d < \deg p$. Prove that there exist unique polynomials q and r over $\mathbf{F}$ (for *quotient* and *remainder*) such that

$$p(x) = d(x)q(x) + r(x).$$

Hint: Mimic the proof of Theorem 3.6. ◇

Inverses

Exercise 3.13 Let $f : \mathbf{R} \to [0, \infty)$ be the squaring function. The function $\sqrt{\ }$ whose value at x is the non-negative square root of x is a branch of f^{-1}. Show that $-\sqrt{\ }$ is another branch of f^{-1}. Find all branches of f^{-1} for this function. (There are many "discontinuous" branches of f^{-1}.) ◇

Exercise 3.14 In this exercise, you will establish a bijection between a bounded interval and $\mathbf{R}$. Define $f : (-1, 1) \to \mathbf{R}$ by $f(x) = x/(1 - x^2)$; see Figure 3.22 for the graph of f.

(a) Set $y = f(x)$ and solve for x.
Suggestion: Multiply by $(1-x^2)$ and rearrange to get the quadratic equation $yx^2 + x - y = 0$. If $y \neq 0$, the quadratic formula applies.

(b) In part (a), you found two formal inverses of f, corresponding to the two signs of the radical in the quadratic formula. You know that f^{-1} must take values in the domain of f, namely in the interval $(-1, 1)$. Which choice of sign is the correct one? What happens when $y = 0$? Identify each choice of sign with a portion of the graph in Figure 3.22.

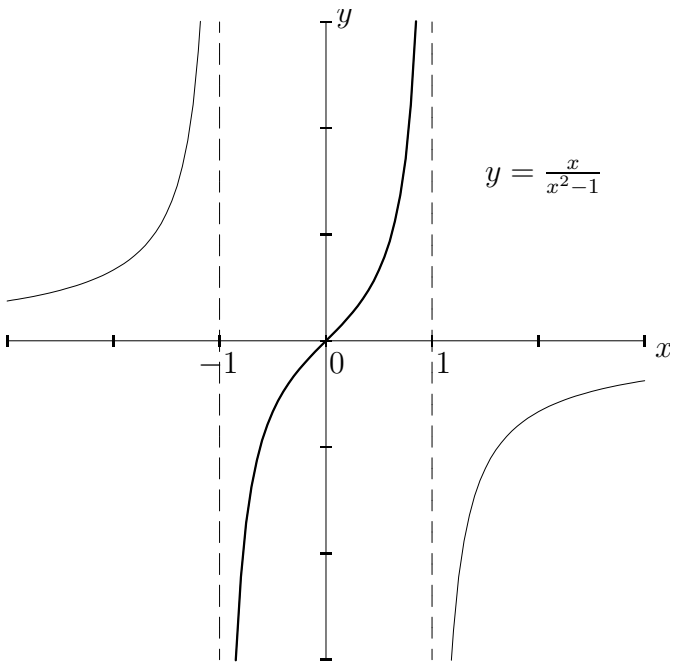


Figure 3.22: A function inducing a bijection from a bounded interval to **R**.

(c) At this stage, a putative formula for f^{-1} has been found. Verify that the formula you found really does give a two-sided inverse of f. That is, verify equations (3.12) and (3.13) directly, or prove by general reasoning that they hold.
Suggestion: If $y > 0$, then $1 + 4y^2 < 1 + 4y + 4y^2 = (1 + 2y)^2$, so

$$0 < \frac{-1 + \sqrt{1 + 4y^2}}{2y} < 1.$$

◇

Symmetries of Functions

Exercise 3.15 Find the even and odd parts of $p(x) = x(x-1)^2$. Find the positive and negative parts of p; write your answer as a piecewise-polynomial function. ◇

Exercise 3.16 Find the even and odd parts of $p(x) = (1-x)^4$.
Hint: Use the binomial theorem to expand p. ◇

Exercise 3.17 Suppose f is even and g is odd. What can you say about their product fg? What if both are odd? Prove all your claims. ◇

Exercise 3.18 Complete the proof of Lemma 3.23 by proving that the set of odd functions is closed under addition and scalar multiplication. ◇

In each of the following exercises, T_ℓ is the translation operator, defined by $T_\ell f(x) = f(x - \ell)$.

Exercise 3.19 Let $f = \chi_{\mathbf{Q}}$ be the characteristic function of $\mathbf{Q}$, and let ℓ be rational. Show that f is ℓ-periodic. ◇

Exercise 3.20 Prove that the set of ℓ-periodic functions is a vector subspace of $\mathscr{F}(\mathbf{R}, \mathbf{R})$. ◇

Exercise 3.21 A function is "ℓ-antiperiodic" if $T_\ell f = -f$. Prove that such a function is 2ℓ-periodic. ◇

Exercise 3.22 Suppose f is ℓ-periodic. Prove that the even and odd parts of f are ℓ-periodic. ◇

Exercise 3.23 Suppose f is 1-periodic, and that g is ℓ-periodic.

(a) Prove that if ℓ is rational, then $f + g$ is periodic.
Suggestion: Write $\ell = p/q$ in lowest terms.

(b) Assume that 1 and ℓ are the *smallest positive periods* of f and g. Prove that if ℓ is irrational, then $f + g$ is *not* periodic.

Part (b) requires serious use of the structure of $\mathbf{Q}$. ◇

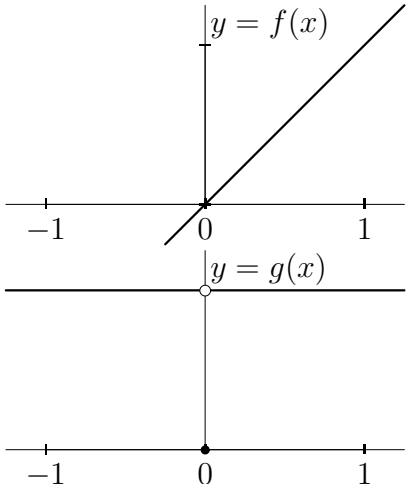
Chapter 4

Limits and Continuity

The concept of "limit" distinguishes analysis (the branch of mathematics encompassing the calculus of differentials) from algebra. The historical motivation and main practical use of limits is to assign meaning to expressions like "$0/0$" or "$0 \cdot \infty$" in a wide variety of situations. As we saw in Chapter 1, describing motion at "an instant of time" leads to difference quotients of the form (distance traveled)/(elapsed time)=$0/0$, while Archimedes' "method of exhaustion" (which allowed him to "dissect" a disk into a rectangle, see Chapter 13) amounts to adding up the areas of infinitely many "infinitely thin" triangles or rectangles, whose total area is formally $0 \cdot \infty$.

A limit is a number that, under certain hypotheses, is assigned to a function f at a point a. However, unlike the function value $f(a)$, which requires consideration of just a single point in the domain, the limit of f at a (if it exists) encodes the behavior of f "near" a, and therefore *cannot* be determined by considering the values of f at only finitely many points! For "continuous" functions (including polynomial, trigonometric, and exponential functions) the "limit of f at a" agrees with $f(a)$. In general there may be a limit at a point where the function value is undefined, or the function value and limit at a point may both exist but be unequal. Before we give any precise definitions, let us consider two simple but illustrative examples:

$$f(x) = x, \qquad g(x) = \begin{cases} 1 & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}, \qquad x \in \mathbf{R}.$$



It is immediately computed that $f(0) = g(0) = 0$; each of these functions vanishes *at* the origin. If instead we try to quantify the behavior *near* the origin, it is believable that (in some sense, which we have not yet made precise) for $|x| \simeq 0$ we have $f(x) \simeq 0$ and $g(x) \simeq 1$. It is a very good philosophical exercise to ponder exactly what might be meant by such an assertion. A few minutes' thought should convince you that consideration of only finitely many function values cannot possibly capture the behavior of f "near" (but not at) a. Instead, to study the behavior of f "near" a we restrict f to arbitrary open intervals about a and consider the image of the restriction. It is therefore in our best interest to develop notation suitable for studying *sets of function values*. The first tool is A notation, which we met in Chapter 2. Two auxiliary notations, "big O" and "little o" (introduced below), will also play prominent roles.

Throughout this chapter, $f : X \to \mathbf{R}$ is a real-valued function *whose domain is a set of real numbers*, usually an interval. We will use the order properties of $\mathbf{R}$ in defining the concept of limit and in proving theorems about limits. It is possible to define limits without an ordering of the domain, but there is additional technical overhead that we wish to avoid.

4.1 Order of Vanishing

In analysis, we are allowed to be a little sloppy; we often don't care if we can solve an equation exactly (whatever this may mean), we only care that a solution is known to exist, and is (say) between 3.14 and 3.1416.

There are calculi (a.k.a., calculational procedures) that allow us to ignore fine details that don't interest us, and concentrate on coarse details that do.

Review of A Notation

Recall that the expression $f = A(\varepsilon)$, read "f is of absolute value at most ε", means that $|f(x)| \leq \varepsilon$ for all x in the domain of f. More generally, if g is a function whose domain contains the domain of f, then to say $f = A(g)$ means $|f(x)| \leq |g(x)|$ for all x. For example, we have $x^2 = A(x)$ on $(-1, 1)$, since $x^2 \leq |x|$ for $-1 < x < 1$.

Our first extension of this terminology allows us to restrict the domain of f by an unspecified amount.

Definition 4.1 If $\varepsilon > 0$, then we say that $f = A(\varepsilon)$ *locally at* a (or *near* a) if there exists some deleted open interval $N_\delta^\times(a)$ on which $f = A(\varepsilon)$. If there exists an $M > 0$ such that $f = A(M)$ near a, then we say f is *locally bounded at* a.

Note that this condition explicitly ignores what happens at a; we might have $|f(a)| > \varepsilon$, or $f(a)$ might not even be defined.

The *smaller* ε is, the *more restrictive* the condition $f = A(\varepsilon)$. For example, if f and g are the functions introduced above, then for each $\varepsilon \geq 1$ we have $g = A(\varepsilon)$ locally at 0, while if $\varepsilon < 1$ it is *not* true that $g = A(\varepsilon)$ near 0. If we ask similar questions about f, we find a possibly surprising answer: If an $\varepsilon > 0$ is given to us, then on the open interval $N_\varepsilon(0) = (-\varepsilon, \varepsilon)$ we have $f = A(\varepsilon)$. In other words, we have $f = A(\varepsilon)$ locally at 0 for *every* $\varepsilon > 0$. Observe carefully that this is not the same thing as saying $f = A(0)$ locally at 0!

There is a potentially confusing point in the last paragraph: In asking whether or not $f = A(\varepsilon)$ locally at a, we are *first* given $\varepsilon > 0$, *then* we choose an interval. Many concepts of analysis similarly depend on one or more "choices" being made, and it is crucially important that the choices be made in an agreed-upon order.

In Chapter 2 we saw informally how A notation is used in calculations. Now we are in a better position to justify these manipulations. If the statements below seem obvious, remember that $f = A(\varepsilon)$ is not an equation, but an abbreviation for "f is of absolute value less than ε".

Proposition 4.2. *If r_1 and r_2 are positive real numbers, then*

$$A(r_1) + A(r_2) = A(r_1 + r2)$$
$$A(r_1) \cdot A(r_2) = A(r_1 r2)$$

In particular, if $x > 0$ *and* $\varepsilon > 0$*, then* $x + A(\varepsilon) = A(x + \varepsilon)$ *and* $xA(\varepsilon) = A(x\varepsilon)$.

Proof. The first assertion is the triangle inequality, but if you are not careful, the inequality can seem to go the wrong way. If $x = A(r_1)$ and $y = A(r_2)$, then $|x + y| \le |x| + |y| \le r_1 + r_2$, which means $x + y = A(r_1 + r_2)$, as claimed. Under the same assumption, $|xy| = |x|\,|y| \le r_1 r_2$, which proves $xy = A(r_1 r_2)$. □

Note carefully that $A(r_1 + r_2) = A(r_1) + A(r_2)$ is *false*. Just because a quantity is no larger than 1 does not mean it is the sum of two quantities each no larger than 1/2.

The expression $x = a + A(\delta)$ means that x is a number of the form a plus a number of absolute value at most δ. This is equivalent to saying $x \in [a - \delta, a + \delta]$. Similarly, if f is a function, then $f = b + A(\varepsilon)$ means the image of f is contained in the interval $[b - \varepsilon, b + \varepsilon]$. Note also that as the number h ranges over an interval about 0, the number $x = a + h$ ranges over an interval about a. Thus $x = a + A(\delta)$ and $x - a = h = A(\delta)$ mean the same thing. These are standard idioms for A notation that you should master.

Example 4.3 The reciprocal function $f(x) = 1/x$, $x \neq 0$, is locally bounded at a for every $a \neq 0$, but is not locally bounded at 0. (It does not matter that 0 is not in the domain, because local boundedness "ignores a".) To see that f is locally bounded at $a \neq 0$, assume first that $a > 0$, and let $\delta = a/2$, Figure 4.1. By the previous paragraph, $x = a + A(\delta)$ means $x \in [\frac{a}{2}, \frac{3a}{2}]$, or that $0 < \frac{a}{2} \le x \le \frac{3a}{2}$. Theorem 2.23 (iv) implies $0 < \frac{2}{3a} \le x \le \frac{2}{a}$, which means $f = A(2/a)$ on the interval of radius $a/2$ centered at a. The case $a < 0$ is similar: take $\delta = -a/2 > 0$. □

In Chapter 3, we saw examples of functions (such as the function that counts the number of "7"s in the decimal expansion of its input) that are not locally bounded at a, no matter which a we are given. This is striking behavior, given that such a function has a well-defined, finite value at each point of $\mathbf{R}$. You should appreciate that local properties are qualitatively different from pointwise properties.

O Notation

A notation allows us to compute with quantities that are not known exactly, but for which we have bounds on the absolute value. Often,

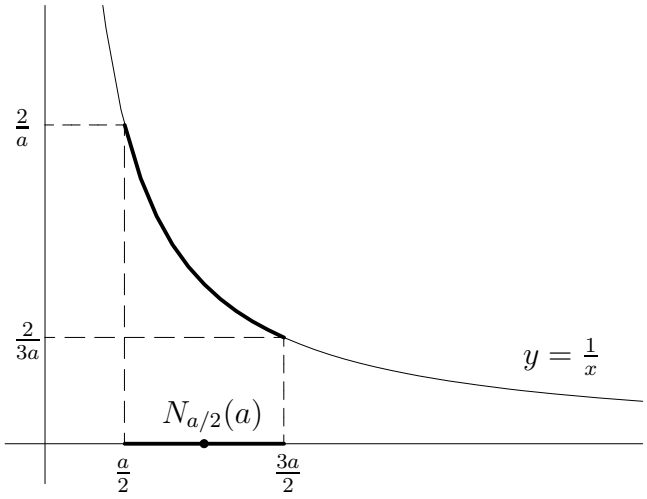


Figure 4.1: Bounding the reciprocal function.

we want to be even more sloppy, and ignore multiplied constants. In this view, we would say that near $x = 0$, x and $10x$ are "roughly the same size", while 1 is definitely larger and x^2 is definitely smaller. If you have used a computer algebra program, you have probably encountered so-called "O notation".

Definition 4.4 Let f and g be real-valued functions whose domains contain some set X. We say that $f = O(g)$ on X (read "f is big-oh of g on X") if there exists a positive real number C such that $|f(x)| \leq C|g(x)|$ for all $x \in X$.

When using O notation, it is important to mention the set X, or at least to keep in mind that there is a restriction on where the inequality holds. For example, we have $O(x^2) = O(x)$ on $[0, 10^{40}]$ (take $C = 10^{40}$), but not on $\mathbf{R}$.

O notation is more symmetric than A notation. Both $O(1) = O(10)$ and $O(10) = O(1)$ are true, for instance. There is an obvious definition of "$f = O(g)$ locally at a", which you should give. We have $x = O(1)$, $10x = O(x)$, and $x^2 = O(x)$ locally at a for each $a \in \mathbf{R}$ (why?). We do not have $x = O(x^2)$ near 0, however.

As with A notation, we can use O notation to calculate with inequalities:

$$\big(1 + O(x)\big)^2 = 1 + 2O(x) + O(x)^2 = 1 + O(x) + O(x^2) \qquad \text{on } \mathbf{R}.$$

In particular, $\big(1 + O(x)\big)^2 = 1 + O(x)$ near 0, and is $O(1)$ near a for

each $a \in \mathbf{R}$. Some important properties of O notation are summarized here.

Proposition 4.5. *Let $h > 0$, and let $k < \ell$ be positive integers. Then*

$$\left.\begin{aligned} O(h^k) + O(h^\ell) &= O(h^k) \\ O(h^k) \cdot O(h^\ell) &= O(h^{k+\ell}) \end{aligned}\right\} \quad \text{near } h = 0$$

If f is a bounded function, then $f \cdot O(h^k) = O(h^k)$.

You will have no difficulty proving these assertions. To see how these properties work in practice, suppose f is a function such that

$$f(x+h) = f(x) + O(h) \text{ near } h = 0, \text{ for all } x \text{ in the domain.} \tag{4.1}$$

Such a function is locally $O(1)$ at each x, since for each x we have

$$\begin{aligned} f(x+h) &= f(x) + O(h) \\ &= f(x) + A(1) = A\big(|f(x)| + 1\big) \\ &= O(1) \qquad \text{near } h = 0, \end{aligned}$$

(remember that x is fixed). Further, if f and g satisfy (4.1), then $f+g$ and fg do as well:

$$\begin{aligned} (f+g)(x+h) &= f(x+h) + g(x+h) \\ &= f(x) + O(h) + g(x) + O(h) \\ &= f(x) + g(x) + O(h) = (f+g)(x) + O(h), \end{aligned}$$

and

$$\begin{aligned} (fg)(x+h) &= \big[f(x) + O(h)\big]\big[g(x) + O(h)\big] \\ &= f(x)g(x) + \big[f(x) + g(x)\big]O(h) + O(h^2) \\ &= (fg)(x) + O(h) \text{ near } 0. \end{aligned}$$

Here are some examples that will be useful in Chapter 8.

Example 4.6 The binomial theorem of Exercise 2.15 implies that

$$(x+h)^n = x^n + nx^{n-1}h + O(h^2) \text{ near } h = 0, \text{ for all } n \in \mathbf{N}. \tag{4.2}$$

The binomial theorem says precisely what the $O(h^2)$ term is equal to, but for many purposes we need only the information furnished by (4.2). For example, we deduce that

$$\frac{(x+h)^n - x^n}{h} = nx^{n-1} + O(h) \text{ near } h = 0.$$

This is useful, because while we cannot set $h = 0$ on the left, we *can* on the right, thereby obtaining an evaluation of $0/0$ is this situation! □

Example 4.7 Suppose f is a function that satisfies the following condition: There exists a real number $f'(0)$ such that

$$f(h) = f(0) + f'(0)h + O(h^2) \text{ near } h = 0.$$

Intuitively, "f is linear up to quadratic terms" at 0. A physicist might write this as $f(h) \simeq f(0) + f'(0)h$ for $h \simeq 0$, but our expression has an explicit, precise interpretation. Now, suppose f and g both satisfy this condition. We immediately calculate that

$$(f+g)(h) = f(0) + g(0) + \big[f'(0) + g'(0)\big]\, h + O(h^2)$$

and

$$\begin{aligned}(fg)(h) &= \big[f(0) + f'(0)\, h + O(h^2)\big]\big[g(0) + g'(0)\, h + O(h^2)\big] \\ &= f(0)g(0) + \big[f'(0)g(0) + f(0)g'(0)\big]\, h + O(h^2),\end{aligned}$$

which proves that $f + g$ and fg also satisfy the condition, and (as a fringe benefit) tells us what $(f+g)'(0)$ and $(fg)'(0)$ are. □

As these examples demonstrate, O notation formalizes "back of the envelope" calculations scientists perform all the time to estimate the predictions of a theory or the outcome of an experiment. More examples are given in the exercises.

o Notation

Our final notational definition looks superficially like O notation, but encapsulates a remarkably subtle, non-trivial property. A single expression in o notation contains *infinitely many* A expressions.

Definition 4.8 Let f be a real-valued function. We say $f = o(1)$ at a if

$$f = A(\varepsilon) \text{ locally at } a \text{ for every } \varepsilon > 0. \tag{4.3}$$

If g is a function that does not vanish on some deleted interval about a, then we say $f = o(g)$ locally at a if $f/g = o(1)$ locally at a.

We saw earlier that the identity function of $\mathbf{R}$ is $o(1)$ at 0; informally, $x = o(1)$ at $x = 0$. Note that x could be replaced by any other letter;

we can (and will) use that fact that $h = o(1)$ at $h = 0$. The condition $f = o(1)$ at a captures the intuition "$f(x)$ can be made arbitrarily small by taking x sufficiently close to a", while $f = o(g)$ at a means that "$f(x)$ is vanishingly small compared to $g(x)$ for x close to a". It is not necessary for f to be defined at a for these assertions to make sense.

The notations o and O stand for *order* (of vanishing). It is convenient to use O and o notations in both theoretical (without variables) and calculational (with variables) settings. The respective notations are slightly different, and you should strive for fluency in both. For instance, the expressions "$f = o(1)$ at x" (no variables) and "$f(x+h) = o(1)$ at $h = 0$" (with variables) mean the same thing. To get a feel for these criteria and their relationships, you should verify the following claims (and, for good measure, translate each into "variables" or "no variables" language, as appropriate):

- If $f = o(1)$ at x, then $f = O(1)$ near x.
- If $f(x+h) = O(h)$ near $h = 0$, then $f(x+h) = o(1)$ at $h = 0$.
- $h^2 = o(h)$ at 0; in fact, $O(h^2) = o(h)$ at 0.
- For each positive integer k, $o(h^k) = O(h^k)$ near 0.

The prototypical vanishing behavior at a is exhibited by the kth power function $g(x) = (x-a)^k$ for $k \in \mathbf{N}$, which we think of as "vanishing to order k" at a. More generally, we say that

$$\text{"}f \text{ vanishes to order } \geq k \text{ at } a\text{" if } f = O\big((x-a)^k\big) \text{ near } a$$
$$\text{"}f \text{ vanishes to order } > k \text{ at } a\text{" if } f = o\big((x-a)^k\big) \text{ near } a$$

We will see shortly that this terminology is natural, and conforms to our intuition about inequalities of integers.

For all functions that one meets in real life (specifically, for "real analytic" functions, which we meet in Chapter 11), the conditions

$$f = O\big((x-a)^{k+1}\big) \quad \text{and} \quad f = o\big((x-a)^k\big)$$

are essentially equivalent. In other words, "most" functions vanish to integer order, except possibly at isolated points. As we saw in Chapter 3, however, there are lots of "perverse" examples of functions, and in general, being $o(1)$ is much less restrictive than being $O(h)$.

The rules for manipulating o expressions in algebraic calculations are similar to those for O notation, but the proofs are more subtle. This is to be expected, since "$f = o(1)$" means "$f = A(\varepsilon)$ for every $\varepsilon > 0$", a condition that involves infinitely many criteria.

Theorem 4.9. *Let f and g be functions defined on some deleted interval about a. If $f = o(1)$ and $g = o(1)$ at a, then $f + g = o(1)$ at a. If $f = o(1)$ and $g = O(1)$ at a, then fg is $o(1)$ at a.*

Informally, the theorem says

$$o(1) + o(1) = o(1), \qquad o(1) \cdot O(1) = o(1).$$

In particular, $c \cdot o(1) = o(1)$ for every real c.

Proof. We are assuming that $f = A(\varepsilon)$ for every $\varepsilon > 0$, and similarly for g. Remember that ε is itself a "variable", standing for an arbitrary positive number. If convenient, we may call it something else, such as $\varepsilon/2$.

We wish to show that $f + g = o(1)$. Let $\varepsilon > 0$. Because $f = o(1)$, there exists an open interval about a on which $f = A(\varepsilon/2)$. Similarly, there exists an open interval on which $g = A(\varepsilon/2)$. On the smaller of these intervals (i.e., on their intersection, see Figure 4.2), we have both $f = A(\varepsilon/2)$ and $g = A(\varepsilon/2)$, and consequently

$$f + g = A(\varepsilon/2) + A(\varepsilon/2) = A(\varepsilon).$$

We have shown that if $\varepsilon > 0$, then there exists an open interval about a on which $f + g = A(\varepsilon)$; this means precisely that $f + g = o(1)$.

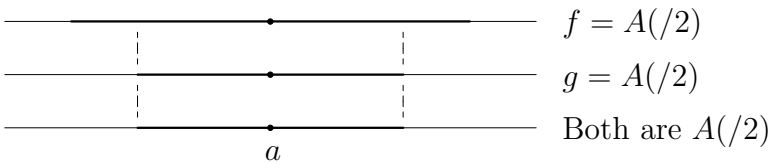


Figure 4.2: Ensuring two conditions on a single open interval.

To establish the second part, begin with the assumption that $g = O(1)$ at a. This means there exists a real number $M > 0$ and an open interval about a such that $|g(x)| \leq M$ for all x in the interval. Now fix $\varepsilon > 0$. Because $f = o(1)$ at a, and because ε/M is a positive real

number, there is an open interval about a on which $f = A(\varepsilon/M)$. As before, consider the smaller open interval; on this interval, we have both $f = A(\varepsilon/M)$ and $g = A(M)$, so

$$fg = A(\varepsilon/M) \cdot A(M) = A(\varepsilon).$$

We have shown that if $\varepsilon > 0$, then $fg = A(\varepsilon)$ on some open interval about a. Since $\varepsilon > 0$ was arbitrary, we have shown that $fg = o(1)$ at a. □

The proof illustrates a standard trick of analysis, depicted in Figure 4.2. If finitely many conditions are given, each holding on some interval about a, then by taking the intersection of these intervals we find a *single* interval on which *all of the conditions hold*. A finite set of intervals at a corresponds to a finite set of positive real numbers (their radii), and the intersection corresponds to the smallest number in the set. By contrast, this trick does not generally apply when there are infinitely many conditions, because the intersection of infinitely many open intervals about a need not be an open interval! To see why, consider the interval $N_{1/n}(a)$ of radius $1/n$ about a. Let us determine which real numbers x are in *all* of these intervals as n ranges over the set of positive integers. Certainly $x = a$ is, by definition. However, if $x \neq a$, then x is not in the intersection: There exists an n such that $1/n < |x - a|$ by the Archimedean property of $\mathbf{R}$, and this inequality means $x \notin N_{1/n}(a)$. Consequently, the intersection of these open intervals is the singleton $\{a\}$. To give a more analytic (less geometric) explanation, recall that an infinite set of positive real numbers always has an *infimum*, but may not have a *minimum*, and the infimum of a set of positive reals can be zero.

4.2 Limits

We are almost ready to introduce the concept of "limit". There is one last technical point that must be raised, regarding domains of functions. So far, we have made no serious assumptions about the domain X of our functions f. However, the language of o has a minor peculiarity. Suppose the domain of f is the single point $\{0\}$. We might still ask, "Is $f = o(1)$ near 1?" The answer is "yes", because in the open interval of radius 1 about 1 there are no points of the domain of f, so vacuously we have "$f = A(\varepsilon)$ near 1 for every $\varepsilon > 0$". We would like to eliminate such

vacuous assertions from our considerations. Also, we want to ignore the value of f at a in defining "limits", since we hope to use the concept of limit even when f is not defined at a. Both of these issues are neatly resolved.

Definition 4.10 Let $X \subset \mathbf{R}$. A point $a \in \mathbf{R}$ is a *limit point of* X if every deleted interval about a contains a point of X. A point of X that is not a limit point of X is an *isolated point*.

The concept of limit point is more subtle than it first appears. For example, whether or not a is a limit point of X is logically independent of whether or not $a \in X$. If a is a limit point of the domain of f, then the condition "$f - f(a) = o(1)$ at a" is non-vacuous. Inversely, if a is *not* a limit point of the domain of f, then the condition *is* vacuous. You should have no trouble verifying a slightly stronger condition, which justifies the term "isolated point" for a point of X that is not a limit point of X:

Lemma 4.11. *Let $X \subset \mathbf{R}$, and let a be a point of $\mathbf{R}$. If a is not a limit point of X, then there exists a deleted interval about a that contains no point of X.*

Example 4.12 Let $X = (0, +\infty)$ be the set of positive real numbers. We claim that the set of limit points of X is the set of non-negative reals. To establish this claim, we will show that every point of $[0, +\infty)$ is a limit point of X, and every point not in $[0, +\infty)$ is not a limit point. The "interesting" point is 0, which is not a point of X, but *is* a limit point of X.

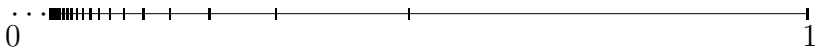
If $a \geq 0$, then for every $\varepsilon > 0$, the point $x = a + \varepsilon/2 > a$ is in the intersection of X with $N_\varepsilon^\times(a)$. This means a is a limit point of X, and establishes the first inclusion. To prove the other direction, suppose $a < 0$; we wish to show a is not a limit point of X. Because the distance from a to X is $|a|$, the deleted interval of radius $\varepsilon = |a|/2 > 0$ is disjoint from X. Because there exists such a deleted interval, the point a is not a limit point of X, as we wished to show. It is a good exercise to translate this geometric argument into the language of inequalities. □

Example 4.13 Our second example is $X = \mathbf{Z}$, the set of integers. This set has no limit points at all! (In particular, a point of X need not be a limit point of X, even if X has infinitely many elements.) If $a \in X$, then the deleted interval of radius 1 contains no point of X, so a is not a limit point of X. If $a \notin X$, then there exists an integer n

such that $n < a < n + 1$ (make sure you can justify this "obvious" assertion using only the axioms of $\mathbf{N}$ and $\mathbf{R}$!), so taking ε to be the smaller of $a - n$ and $n + 1 - a$, we see that the deleted interval of radius $\varepsilon > 0$ contains no point of X, and again a is not a limit point of X. □

Example 4.14 Our third example is $X = \mathbf{Q}$, the set of rational numbers. Recall (Theorem 2.33) that every open interval of reals contains a rational number. It follows easily that *every* real number is a limit point of $\mathbf{Q}$: If $a \in \mathbf{R}$ and $\varepsilon > 0$, then the open interval $(a, a + \varepsilon) \subset N_\varepsilon^\times(a)$ contains a point of $\mathbf{Q}$. □

As a test of your understanding, determine (with proof!) the set of limit points of the interval $(0, 1)$, and of the set $X = \{1/n \mid n > 0\}$:



Most of the points of X are not depicted here; at the left edge the tick marks that represent them run together. It may help to break the problem into cases. First treat the cases $a < 0$ and $a > 1$; then consider the points of X itself, keeping in mind what X looks like as you "zoom in" close to 0; next consider points $0 < a < 1$ that are not elements of X; and finally consider 0. If you feel that you are landing a plane in fog, remember that the definitions are your radar.

Finally we can introduce limits. We want to say a real number ℓ is a "limit" of f at a if $f = \ell + o(1)$ at a. The fine print is that we wish to avoid a vacuous statement (so we require that a is a limit point of the domain of f), and we wish to ignore the value of f at a (hence we restrict to a deleted interval). The completely unraveled definition is the "ε-δ criterion" well-known to all students of analysis. We have packed it into o notation in order to clarify the conceptual content.

Definition 4.15 Let $f : X \to \mathbf{R}$ be a function, and let a be a limit point of X. The real number ℓ is said to be a *limit of f at a* if $f = \ell + o(1)$ at a. In this situation, we write $\ell = \lim(f, a)$ or $\ell = \lim_{x \to a} f(x)$.

If "$\lim(f, a) = \ell$" is false for every real number ℓ, then we say that the limit of f at a *does not exist*. The way we have set up the definition, it does not make sense to ask whether or not $\lim(f, a)$ exists unless a is a limit point of the domain of f.

Limits are often explained informally by saying " 'ℓ is a limit of f at a' means that 'the function values $f(x)$ can be made arbitrarily close to ℓ provided x is sufficiently close to a'." If you have studied limits elsewhere, you have likely see the following definition:

> The real number ℓ is a *limit of f at a* if, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $0 < |x - a| < \delta$ implies $|f(x) - \ell| < \varepsilon$.

It is straightforward to see that this condition is equivalent to Definition 4.15. The phrase "$0 < |x - a| < \delta$ implies $|f(x) - \ell| < \varepsilon$" means that $f = \ell + A(\varepsilon)$ on the deleted δ-interval about a. The clause "for every $\varepsilon > 0$, there exists $\delta > 0$" means, in this context, that "for every $\varepsilon > 0$, there is a deleted interval…". The definition above therefore means that $f = \ell + o(1)$ at a.

The expression "$\lim(f, a) = \ell$" looks like an equation, but you should be wary of treating it as such; remember that the expression "$f = o(g)$" is not an *equation*, but an *abbreviation*. Explicitly, the problem is that more than one number could conceivably arise as a limit of f at a. The first order of business is to dispel this worry:

Lemma 4.16. *If ℓ and m are* constant *functions on some deleted interval about a, and if $\ell - m = o(1)$ at a, then $\ell = m$. In words, if two real numbers are arbitrarily close, then they are equal.*

Proof. Strangely, the hypothesis consists of *infinitely many* statements, namely that $\ell - m = A(\varepsilon)$ at a for each $\varepsilon > 0$, but *no finite number* of these assumptions imply the conclusion! Instead, consider the contrapositive: If $\ell \neq m$, then there exists $\varepsilon > 0$ such that $\ell - m$ is *not* $A(\varepsilon)$. With a moment's thought, this is obvious: If $\ell - m \neq 0$, then take $\varepsilon = |\ell - m|/2 > 0$. By hypothesis, we have $|\ell - m| = 2\varepsilon$ on every deleted interval about a, and $2\varepsilon > \varepsilon$ because $\varepsilon > 0$. □

Theorem 4.17. *Let f be a function. If $\lim(f, a) = \ell$ and $\lim(f, a) = m$, then $\ell = m$.*

Informally, f has at most one limit at a, and it makes sense to speak of *the* limit of f at a (provided we agree the limit may not exist). Theorem 4.17 also justifies treating the expression "$\lim(f, a) = \ell$" as an equation of real numbers when the limit exists, and we will do so from now on.

Proof. Suppose $f = \ell + o(1)$ and $f = m + o(1)$ at a. This means that $f - \ell = o(1)$ and $m - f = o(1)$ at a. Adding, we have $\ell - m = o(1)$ at a, which implies $\ell = m$ by Lemma 4.16. □

The notations

$$\lim_{x\to a} f(x) = \ell \quad \text{and} \quad \lim(f, a) = \ell \tag{4.4}$$

are read, "The limit of $f(x)$ as x approaches a is equal to ℓ" and "the limit of f at a is ℓ". The latter is more concise in abstract situations as it avoids introduction of the spurious "x." It should be emphasized that while "$\lim_{x\to a} x^2$" is permissible (because "$x \to a$" makes it clear that "x^2" is the function value at x), the expression "$\lim(x^2, a)$" is ambiguous because "x^2" does not define a function. The two likeliest meanings are $\lim_{x\to a} x^2 = a^2$ and $\lim_{t\to a} x^2 = x^2$.

In order to use limits, we need the "usual tools": theorems that give general properties of limits, examples of functions that do and do not have limits, and easy calculational methods for finding and working with limits.

Example 4.18 Two limits are immediate: If c is a real number, then $\lim_{x\to a} c = c$, and $\lim_{x\to a} x = a$ for all a. In o notation, $c = c + o(1)$ and $x = a + o(1)$ near $x = a$. The first is obvious, because $0 = o(1)$, and the second is clear because $h = o(1)$ at $h = 0$. □

Theorem 4.19. *Let f and g be functions having the same domain X. If $\lim(f, a) = \ell$ and $\lim(g, a) = m$, then $\lim(f + g, a)$ and $\lim(fg, a)$ exist, and are equal to $\ell + m$ and ℓm respectively. If in addition $m \neq 0$, then $\lim(f/g, a)$ exists, and is equal to ℓ/m.*

Proof. (Sums) The hypothesis is that $f = \ell + o(1)$ and $g = m + o(1)$ at a; by Theorem 4.9, we have $f + g = \ell + m + o(1)$ at a, which means $\lim(f + g, a) = \ell + m = \lim(f, a) + \lim(g, a)$.

(Products) Under the same hypotheses, we have

$$\begin{aligned} fg &= \big(\ell + o(1)\big)\big(m + o(1)\big) && \text{by hypothesis} \\ &= \ell m + (\ell + m)o(1) + o(1) \cdot o(1) \\ &= \ell m + o(1) && \text{Theorem 4.9.} \end{aligned}$$

This means $\lim(fg, a) = \ell m = \lim(f, a) \cdot \lim(g, a)$.

(Quotients) This is a little more involved, but the only new ingredient is the technique of Example 4.3, see Figure 4.1. Assume first that $m > 0$, and let $\varepsilon = m/2 > 0$. Because $g = m + o(1)$, there is a deleted interval about a on which $g = m + A(\varepsilon)$, which by choice of ε means

$$\frac{m}{2} < g < \frac{3m}{2}, \quad \text{or} \quad \frac{2}{3m} < \frac{1}{g} < \frac{2}{m}.$$

In other words $1/g = O(1)$ near a. Direct calculation gives

$$\begin{aligned}\frac{1}{g} - \frac{1}{m} &= \frac{m-g}{gm} \\ &= -(g-m) \cdot \frac{1}{g} \cdot \frac{1}{m} \\ &= o(1) \cdot O(1) \cdot O(1) = o(1),\end{aligned}$$

or $1/g = (1/m) + o(1)$ at a. Multiplying by $f = \ell + o(1)$, we have $f/g = \ell/m + o(1)$, as was to be shown. □

Corollary 4.20. *Let p and q be polynomials with no common factor, and let $f = p/q$ be the corresponding rational function. If $q(a) \neq 0$, then $\lim(f, a) = p(a)/q(a)$. In particular, if $p : \mathbf{R} \to \mathbf{R}$ is a polynomial function, then $\lim(p, a) = p(a)$ for all $a \in \mathbf{R}$.*

In words, limits of rational functions are obtained by evaluation. For example,

$$\lim_{x \to a} \frac{1 + x + x^5}{4 - x^2} = \frac{1 + a + a^5}{4 - a^2}$$

if $a \neq \pm 2$.

Proof. Every monomial function $x \mapsto a_n\, x^n$ is a product of a constant and n copies of the identity function, and therefore satisfies the conclusion by the "products" part of Theorem 4.19. The same is true of polynomials by the "sums" part of the theorem, and for rational functions by the "quotients" part. To formalize this argument completely, you would do several arguments by mathematical induction. It is worth proving the assertion for monomials carefully, to get a feel for what is involved. □

Corollary 4.21. *Suppose $\lim(f, a) = \ell$ exists, but that g has no limit at a. Then $f + g$ has no limit at a, and if $\ell \neq 0$, then fg has no limit at a.*

Proof. Let $h = f + g$; the theorem asserts that if h has a limit at a, then so does $g = h + (-f)$. This is the contrapositive of the corollary. The claim about fg is Exercise 4.5. □

The fact that limits and pointwise evaluation are the same thing for rational functions might suggest that the concepts are always the same. Philosophically, almost the exact opposite is true. Theorem 4.22, the

locality principle, makes precise the assertion that limits cannot be determined by looking at finitely many function values. This "blindness" to finite amounts of data seems paradoxical, but there is no logical inconsistency. The real lesson is that the number line is more complicated than first meets the eye.

Theorem 4.22. *Let f and g be real-valued functions whose domains contain some deleted interval about $a \in \mathbf{R}$, and assume that $f(x) = g(x)$ except possibly at finitely many x. Then $\lim(f, a) = \lim(g, a)$ in the sense that either both limits exist and are equal, or else neither limit exists.*

Proof. Let $\{b_1, \ldots, b_n\}$ be the list of distinct points at which $f(x) \neq g(x)$. We may as well assume none of the b_i is a, since we are interested only in behavior of f and g on deleted intervals about a. Let $\delta > 0$ be the minimum of $|b_i - a|$ for $1 \leq i \leq n$, namely the distance from a to the closest of the b_i. (It is here that finiteness is used; if there were infinitely many b_i, we could not guarantee $\delta > 0$.) On the deleted interval $N_\delta^\times(a)$, f and g are *the same function*, so for the purposes of Definition 4.15, $f = g$. □

Another useful, fundamental property is that "limits respect $\leq$". You should note carefully that information about limits at a single point tells you something about functions on an open interval.

Theorem 4.23. *Suppose f and g have the same domain, that $\lim(f, a)$ and $\lim(g, a)$ exist, and that $f(x) \leq g(x)$ for all x in some deleted interval about a. Then $\lim(f, a) \leq \lim(g, a)$.*

Proof. Though the statement given in the theorem arises frequently in applications, the contrapositive is more natural to prove: If $\lim(f, a) > \lim(g, a)$, then there is a deleted interval about a on which $f > g$.

Consider the function $h = f - g$, which has limit

$$\ell := \lim(f, a) - \lim(g, a) > 0$$

by Theorem 4.19. We will show there exists a deleted neighborhood of a on which $h > 0$. Let $\varepsilon = \ell/2 > 0$. Because $h = \ell + A(\varepsilon)$ locally at a, there exists a deleted interval about a on which $h > \ell - \varepsilon = \ell/2 > 0$. This already proves the theorem, but it is worth mentioning that not merely is $h > 0$ near a, but h has a *positive lower bound*, or is *bounded away from zero*. This extra piece of information is often useful. □

If the hypothesis is strengthened to "$f(x) < g(x)$ for all x in some deleted interval," it *does not follow* in general that $\ell < m$ (though of course $\ell \leq m$, by the theorem). You should examine the proof to see why not, and find a pair of functions $f < g$ for which $\lim(f, 0) = \lim(g, 0)$.

A related result is the so-called "squeeze theorem" (or "pinching theorem"). Several interesting limits are evaluated by finding simple upper and lower bounds and applying the squeeze theorem.

Theorem 4.24. *Suppose $f \leq h \leq g$ on some deleted interval about a, and that $\lim(f, a)$ and $\lim(g, a)$ both exist and are equal to ℓ. Then $\lim(h, a)$ exists and is equal to ℓ.*

Proof. Fix $\varepsilon > 0$, and choose $\delta > 0$ so that $f = \ell + A(\varepsilon)$ and $g = \ell + A(\varepsilon)$ on the deleted δ-interval about a. Combining this with the hypothesis that $f \leq h \leq g$ locally at a, we have $-\varepsilon \leq f - \ell \leq h - \ell \leq g - \ell < \varepsilon$ on $N_\delta^\times(a)$, so $h = \ell + A(\varepsilon)$ locally at a. Since $\varepsilon > 0$ was arbitrary, $\lim(h, a) = \ell$. □

Example 4.25 Let $\operatorname{sgn} : \mathbf{R} \to \mathbf{R}$ be the signum function. The locality principle implies that if $a \neq 0$, then

$$\lim_{x \to a} \operatorname{sgn}(x) = \operatorname{sgn}(a) = \frac{a}{|a|} = \begin{cases} 1 & \text{if } a > 0, \\ -1 & \text{if } a < 0. \end{cases}$$

Near 0, however, sgn takes the values 1 and -1; these do not lie in any interval of length less than 2, so if $\varepsilon \leq 1$ the condition $\operatorname{sgn} = \ell + A(\varepsilon)$ is *false for every real* ℓ. This means that $\lim(\operatorname{sgn}, 0)$ does not exist. □

Example 4.26 Here are two relatively involved examples. The first has no limit at a for every point a in the domain. The second has a limit at every point while at first glance it seems to have a limit nowhere.

Consider the function $\chi_{\mathbf{Q}} : \mathbf{R} \to \mathbf{R}$, the characteristic function of the set of rational numbers, and fix a real number a. In every deleted interval about a, there are both rational and non-rational real numbers, so the function $\chi_{\mathbf{Q}}$ takes both values 0 and 1. No matter how the potential limit ℓ is chosen, if $\varepsilon < 1/2$ then we do not have $\chi_{\mathbf{Q}} = \ell + A(\varepsilon)$ locally at a, because the "target" interval $(\ell - \varepsilon, \ell + \varepsilon)$ has length < 1. This means $\chi_{\mathbf{Q}}$ has no limit at a, for every real number a.

The second example is the "denominator" function f of Example 3.11. An enlarged portion of the graph is depicted in Proposition 4.28 below. This function bears a certain resemblance to the characteristic function of $\mathbf{Q}$, though the function values are smaller when the denominator is smaller. In order to study the limit behavior of f, it is convenient to consider the set

$$\mathbf{Q}(N) := \{p/q \in \mathbf{Q} : |q| \leq N\} = \bigcup_{q=1}^{N} \tfrac{1}{q}\mathbf{Z}$$

of rational numbers whose denominator is at most N, and to write $\mathbf{Q}$ as the union of these sets as N ranges over the positive integers.

We showed in Example 4.13 that the set $\mathbf{Z}$ of integers has no limit point, and an obvious modification of the argument proves that similarly, the set $\frac{1}{q}\mathbf{Z}$ has no limit point. The following remark shows that the set $\mathbf{Q}(N)$ itself has no limits point.

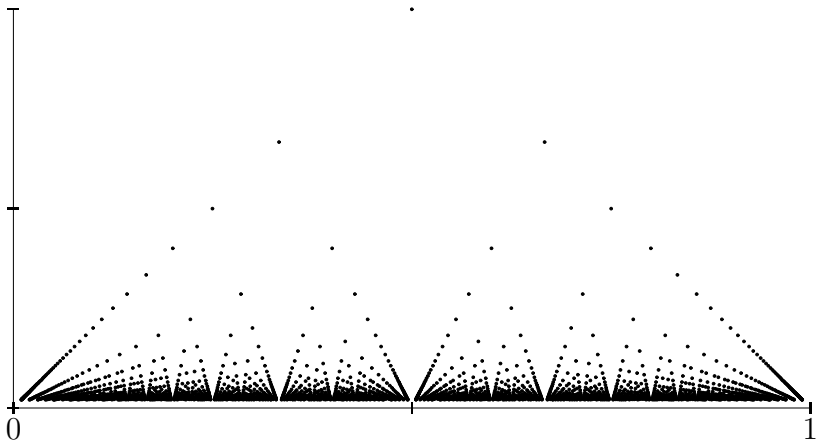
Lemma 4.27. *Let $A_1, \ldots, A_N$ be subsets of $\mathbf{R}$ that have no real limit point. Then their union has no limit point.*

Proof. Let a be a point of $\mathbf{R}$. Because A_1 has no limit point, there exists a deleted interval I_1 about a that contains no point of A_1, by Lemma 4.11. Arguing similarly, we have finitely many deleted intervals $I_2, \ldots, I_n$, such that I_k contains no point of A_k. The intersection of the I_k is a non-empty deleted interval about a that contains no point of $A_1 \cup \cdots \cup A_n$. In particular a is not a limit point of $A_1 \cup \cdots \cup A_n$. □

Proposition 4.28. *Define $f : \mathbf{R} \to \mathbf{R}$ by*

$$f(x) = \begin{cases} 1/q & \text{if } x = p/q \text{ in lowest terms,} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

Then $\lim(f, a) = 0$ for every $a \in \mathbf{R}$.



Proof. Recall that $\mathbf{Q}$ is the union over $N \geq 1$ of the sets $\mathbf{Q}(N)$. By definition of f, elements of $\mathbf{Q}(N)$ are exactly the real points x for which $|f(x)| \geq 1/N$. This fact accounts for the narrow band near the horizontal axis in the figure; only points with $N < 100$ are shown.

Let a be an arbitrary real number. Fix $\varepsilon > 0$, and choose $N \in \mathbf{N}$ so that $1/N < \varepsilon$. By Lemma 4.27, there exists a deleted interval about a that contains no point of $\mathbf{Q}(N)$. Since this interval omits all points x for which $f(x) \geq 1/N$, we have $f = A(\varepsilon)$ locally at a. But ε was arbitrary, so we have shown that $f = o(1)$ at a for every a. □

Limits are unmistakably "new" information about a function that cannot be seen by considering individual function values. The previous example uses the full power of the limit definition, and shows how the definition can depart from intuition. Though the denominator function is non-zero at infinitely many points, its limit exists and is equal to zero at *every* real number a. Proposition 4.28 can be stated as a sort of approximation principle: If a real number a is approximated by rational numbers $x \neq a$, then the denominators must grow without bound. This is true even if $a \in \mathbf{Q}$! □

The Limit Game

There is a game-theoretic interpretation of limits that many people find helpful, presumably because the human brain is better wired to understand competition than existential quantifiers. Imagine a two-player game, with participants Player ε ("your opponent") and Player δ ("you"). A referee specifies in advance a function f, a point a that is a limit point of the domain of f, and a real number ℓ (the putative

limit). Player ε goes first, choosing a positive number ε. This number is a "tolerance" on function values, and specifies the radius of a target centered at ℓ. To "meet" the tolerance (or hit the target) for a point x in the domain of f means that $|f(x) - \ell| < \varepsilon$; making ε smaller makes the target more difficult to hit.

Now it is your turn: You want to choose a "launching radius" ($\delta > 0$) so that every "shot" x that originates in the deleted δ-interval about a hits the target. Clearly, a smaller choice of launching radius does not make your accuracy any worse. If you succeed, then you win the round, and we say that $f = \ell + A(\varepsilon)$ locally at a; otherwise Player ε wins.

The equation $\lim(f, a) = \ell$ means that "You have a winning strategy against a perfect player": No matter how small Player ε makes the target (remember, its size is positive), you can win the game. The distinction between winning one round and having a winning strategy against a perfect player is the distinction between having $f = \ell + A(\varepsilon)$ for a *particular* $\varepsilon > 0$, and having $f = \ell + A(\varepsilon)$ for *arbitrary* $\varepsilon > 0$.

As mentioned previously, it is crucial that ε plays *first*. Whether or not you have a winning strategy is determined by the function f, the location a, and the putative limit ℓ; it does not depend on the choice of ε. If Player ε blunders, then you may win a round even if "$\lim(f, a) = \ell$" is false, compare Example 4.25. As risk of being overly repetitive, winning a round is not as good as having a *winning strategy against a perfect player*.

If you find the game-theoretic interpretation helpful, you may wish also to consider the following variant: The game starts exactly as before, with the referee's choice of f, a, and ℓ. Your opponent chooses $\varepsilon > 0$, and you choose $\delta > 0$, but now your opponent chooses an x in the domain of f that satisfies $0 < |x - a| < \delta$. If $|f(x) - \ell| < \varepsilon$ then you win the round, and otherwise you lose. Again, "$\lim(f, a) = \ell$" is equivalent to your having a winning strategy against a perfect player.

There are other mathematically interesting "limit games" that arise from small changes in the rules. For example, the referee might specify $\ell = f(a)$; or might not specify the number a, requiring you to win simultaneously for all $a \in A$ with the choice $\ell = f(a)$; or might require that you win for all $a \in A$ with $\ell = f(a)$ and *with a single choice of* δ. We will meet these and other limit games in due course.

The "limit game" is idealized in an important way: A putative limit ℓ is specified in advance. In an actual situation, little or nothing is known about ℓ. The definition will only say whether or not the limit is equal to ℓ, not whether or not the limit actually exists. Almost nothing is

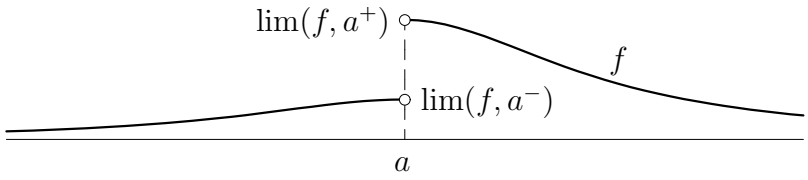


Figure 4.3: One-sided limits of f at a.

learned if the attempt to show $\lim(f, a) = \ell$ fails; the referee supplied the "wrong" number ℓ, but there may be no "right" number. In the game analogy, if you think there is a target centered at ℓ, and you shoot accurately but miss, you cannot deduce that there is no target, only that you aimed at the wrong location. There are procedures for proving *existence* of a limit without actually knowing what ℓ is; such a theorem is a kind of "radar" that detects a target without locating it. This knowledge alone can be useful for a couple of reasons. First, there are theorems for finding a limit if it is known that one exists (and additional hypotheses are satisfied); perhaps even more importantly, it means that new functions can be defined as limits extracted from given functions. Derivatives, integrals, and power series—three pillars of calculus—are of this type.

One-Sided Limits

The limit of f at a takes into account the behavior of f on deleted intervals about a. The deleted interval $N_r^\times(a)$ is the union of two open intervals, $(a - r, a)$ and $(a, a + r)$, and it is sometimes desirable to study f on each interval separately.

If the function f is defined on the interval $(a, a+r)$ for some $r > 0$, and if (after restriction to this open interval) $f = \ell + o(1)$ near a, then we say the *limit from the right of f at a* (or "from above") exists, and write

$$\lim(f, a^+) = \ell \qquad \text{or} \qquad \lim_{x \to a^+} f(x) = \ell \qquad \text{or} \qquad \lim_{x \searrow a} f(x) = \ell$$

You should translate this condition into an ε-δ game, and should formulate the analogous definition for limits from the left (i.e., "from below") at a. The notation for limits from the left is as expected:

$$\lim(f, a^-) = \ell \qquad \text{or} \qquad \lim_{x \to a^-} f(x) = \ell \qquad \text{or} \qquad \lim_{x \nearrow a} f(x) = \ell.$$

If $\lim(f, a) = \ell$, then both one-sided limits exist and are equal to ℓ. You should have no trouble showing that, conversely, if both one-sided limits of f at a exist and are equal to ℓ, then $\lim(f, a) = \ell$. Despite this close correspondence, one-sided limits arise naturally in important situations, especially when we wish to prove existence of a limit without having specific information about f, see Theorem 4.30.

Example 4.29 Let $\operatorname{sgn} : \mathbf{R} \to \mathbf{R}$ be the signum function. Because $\operatorname{sgn} x = 1$ for all $x > 0$, we have $\lim(\operatorname{sgn}, 0^+) = 1$; similarly, $\lim(\operatorname{sgn}, 0^-) = -1$. Combined with the observations of the previous paragraph, we deduce once again that $\lim(\operatorname{sgn}, 0)$ does not exist, compare Example 4.25. □

Limits of Monotone Functions

Monotone functions have simple limit behavior; this is disguised in the completeness property of $\mathbf{R}$ and accounts for the crucial role of completeness in analysis. Theorem 4.30 is stated for non-decreasing functions (increasing functions in particular). There is an obvious analogue for non-increasing functions. This result is clear evidence that the concept of "supremum" is the correct generalization of "maximum" for bounded sets with infinitely many elements.

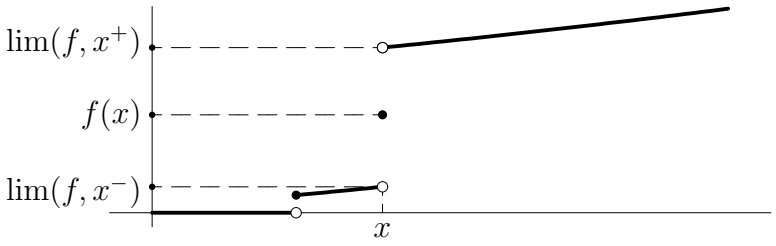


Figure 4.4: One-sided limits of a monotone function.

Theorem 4.30. *Let $A = (a, b)$ be an open interval, and let $f : A \to \mathbf{R}$ be a non-decreasing function. Then the one-sided limits $\lim(f, x^{\pm})$ exist for all $x \in A$.*

Proof. To show a function has a limit, it is necessary to have a candidate limit. For a non-decreasing function, the inclination is to guess that the limit from below at x is the "maximum" of all function values $f(y)$

with $y < x$. (We don't want to allow $y = x$ since "limits at x do not see the function value at x.") Unfortunately, the set $\{f(y) \mid y < x\}$ generally has no maximum; all we know is that the set is non-empty and bounded above (by $f(x)$ itself, since f is non-decreasing). However, these conditions are exactly enough to imply the set has a supremum, and we are led to *guess* that if $f : (a, b) \to \mathbf{R}$ is non-decreasing, then

$$\begin{aligned} \ell^- &:= \lim(f, x^-) = \sup\{f(y) \mid a < y < x\}, \\ \ell^+ &:= \lim(f, x^+) = \inf\{f(y) \mid x < y < b\}. \end{aligned} \tag{4.5}$$

This guess is not only correct, but easy to establish from the definitions. Fix $\varepsilon > 0$; by definition of supremum, $\ell^- - \varepsilon$ is *not* an upper bound of $\{f(y) \mid a < y < x\}$, so there is a $z < x$ with $\ell^- - \varepsilon < f(z) \le \ell^-$. Let $\delta = x - z$; then $\delta > 0$, and because f is non-decreasing, $z < y < x$ implies $f(z) \le f(y) \le \ell^-$. But $z = x - \delta$, so the previous implication can be written

$$x - \delta < y < x \implies \ell^- - \varepsilon < f(y) \le \ell^-,$$

which trivially implies $|f(y) - \ell^-| < \varepsilon$. This means the limit from below at x is ℓ^-, as claimed. The other one-sided limit is established by an entirely similar argument, see Exercise 4.19. □

Limits at Infinity

Recall that the *extended real numbers* are obtained by adjoining two non-real points, $+\infty$ and $-\infty$, to $\mathbf{R}$ and extending the order relation so that $-\infty < x < +\infty$ for all real x. The symbols $\pm\infty$ do not represent real numbers, and for now arithmetic operations with them are undefined. Intuitively, $-\infty$ and $+\infty$ are the "left and right endpoints of $\mathbf{R}$." We have so far considered limits of functions at (finite) limits points. For many practical and theoretical applications, we wish to define and use limits at $\pm\infty$. For example, long-time behavior of a physical system is often modeled by a limit at $+\infty$.

The analogue of a deleted interval at $+\infty$ is an interval of the form $(R, +\infty)$ for some $R \in \mathbf{R}$; the interval shrinks as R gets *larger*, that is, "closer to $+\infty$." Similarly, a deleted interval about $-\infty$ is an interval of the form $(-\infty, R)$. Everything in this section has an analogue at $-\infty$, but to simplify the exposition, we shall explicitly mention only $+\infty$.

The definitions and theorems for limits carry over to this new situation without change. Let f be a real-valued function whose domain X is

a subset of $\mathbf{R}$. We say that $+\infty$ is a *limit point* of X if X is not bounded above; intuitively, X contains points arbitrarily close to $+\infty$. In low-level detail, $+\infty$ is a limit point of X iff for every $R \in \mathbf{R}$, there exists a point $x \in X$ with $x > R$. It makes sense to say $f = A(\varepsilon)$ near $+\infty$; this means that if we restrict f to a deleted interval about $+\infty$, the image is contained in $[-\varepsilon, \varepsilon]$, just as in the finite case. Similarly, we may say $f(h) = O(1/h)$ near $+\infty$, or $f = o(1)$ at $+\infty$. The precise interpretations are left to you.

Definition 4.31 Let $f : X \to \mathbf{R}$ be a function whose domain is not bounded above. A real number ℓ is a *limit at* $+\infty$ of f if, for every $\varepsilon > 0$, $f = \ell + o(1)$ at $+\infty$.

As for limits at finite points, limits at $+\infty$ are unique (if they exist), so it is permissible to treat the expressions

$$\lim_{x \to +\infty} f(x) = \ell \text{ and } \lim(f, +\infty) = \ell$$

as equations of real numbers. We will take for granted that theorems established for limits at finite points carry over to limits at $\pm\infty$.

Example 4.32 Every constant function has the obvious limit at $+\infty$; the identity function $I(x) = x$ has no limit at $+\infty$. (Arguably the limit is $+\infty$, and with an appropriate definition this is a theorem. However, for a function to have infinite limit is a special case of having no real limit.)

A limit at $+\infty$ fails to exist for a different reason when $f : \mathbf{R} \to \mathbf{R}$ is a non-constant *periodic* function: The function "cannot decide what value to approach". Formally, let x and y be points with $f(x) \neq f(y)$. On every interval $(R, +\infty)$, f achieves the values $f(x)$ and $f(y)$, but there is no number ℓ such that *every* interval about ℓ contains both $f(x)$ and $f(y)$. □

Example 4.33 Recall that a (real) sequence is a function $a : \mathbf{N} \to \mathbf{R}$, usually denoted $(a_k)_{k=0}^{\infty}$. A sequence (a_k) *converges* to $\ell \in \mathbf{R}$ if $\lim_{k \to \infty} a_k = \ell$. This condition is discussed in detail in Section 4.4. □

The reciprocal function $f(x) = 1/x$ for $x \in (0, +\infty)$ is easily seen to approach 0 as $x \to +\infty$: If $\varepsilon > 0$, then $f = A(\varepsilon)$ on the interval $(1/\varepsilon, +\infty)$. More generally, if $k \geq 1$ is an integer, then $\lim_{x \to +\infty} 1/x^k = 0$:

$$\lim_{x \to +\infty} \frac{1}{x^k} = \left(\lim_{x \to +\infty} \frac{1}{x} \right)^k = 0^k = 0,$$

the first equality being "the limit of a product is the product of the limits." This equality can be established formally by mathematical induction. Alternatively, you can use the squeeze theorem.

Using these preliminary results, we can find the limits of rational functions at $+\infty$. In the next example, dividing the numerator and denominator by x^4 and using Theorem 4.19 gives

$$\lim_{x\to+\infty} \frac{1-2x+3x^2}{x+x^4} = \lim_{x\to+\infty} \frac{(1/x^4)-(2/x^3)+(3/x^2)}{(1/x^3)+1} = \frac{0-2\cdot 0+3\cdot 0}{0+1},$$

so the limit is 0. In general, the limit exists iff the numerator has degree no larger than the denominator, and the limit is non-zero iff the numerator and denominator have the same degree:

Theorem 4.34. *Let p and q be polynomials with no common factor, say*

$$p(x) = \sum_{k=0}^{n} a_k\, x^k, \qquad q(x) = \sum_{k=0}^{m} b_k\, x^k, \qquad a_n,\ b_m \neq 0.$$

If $n < m$, then $\lim(p/q, +\infty) = 0$, while if $n = m$ the limit exists and is equal to a_n/b_n. If $n > m$ the limit does not exist.

Proof. Intuitively, the largest-degree terms are the only ones that matter, and the proof essentially exploits this idea. First, the denominator has at most finitely many zeros, so the quotient is defined on some interval $(R, +\infty)$. Divide the numerator and denominator by their highest respective powers of x; the resulting expression is

$$\frac{p(x)}{q(x)} = x^{n-m} \cdot \left(\sum_{k=0}^{n} a_k\, x^{k-n}\right) \Big/ \left(\sum_{k=0}^{m} b_k\, x^{k-m}\right).$$

The terms in parentheses individually have limits because they are sums of monomials that have limits; the numerator approaches a_n, the denominator approaches b_m, so their quotient approaches a_n/b_m by Theorem 4.19.

If $n \leq m$, then the claim is immediate since the "leftover" term x^{n-m} approaches 0 if $n < m$ and is identically 1 if $n = m$. If $n > m$, then this term has no limit at $+\infty$, so the quotient p/q has no limit by Corollary 4.21. □

If the domain of f contains some interval of the form $(R, +\infty)$, and if $\lim(f, +\infty) = \ell$, then the line $y = \ell$ is called a *horizontal asymptote*

of the graph of f. Theorem 4.34 says the graph of a rational function p/q has a horizontal asymptote provided the degree of the numerator does not exceed the degree of the denominator.

Infinite Limits

There are two contexts in which "infinity" may be treated as a "limit." In the first of these, two points ($+\infty$ and $-\infty$) are appended to $\mathbf{R}$, and we distinguish between large positive function values and large negative[1] function values. In the second context, only a *single* point, called ∞, is appended to $\mathbf{R}$, and we may no longer speak of order relations involving ∞. Each technique is useful in different situations, and each will be studied in turn. To avoid a certain amount of repetition, all functions in this section are assumed to have unbounded domain.

Extended Real Limits

The definition of finite limits makes sense because we know what is meant by the condition $f = \ell + A(\varepsilon)$. If we replace ℓ by $+\infty$, this is no longer true, because we cannot perform algebraic operations with $+\infty$. In order to proceed, we must reformulate the definition of finite limits in a way that makes sense "when $\ell = +\infty$". To say that $\lim(f, a) = \ell$ means that for every interval I about ℓ there exists a deleted interval about a such that the image of the restriction of f is contained in I. This looks good, because we know what is meant by an interval about $+\infty$, and the condition does not mention algebraic operations with ℓ.

Definition 4.35 Let f be a function. We say *the limit of f at a is $+\infty$* (or "the limit of $f(x)$ as x approaches a is $+\infty$") if, for every $R \in \mathbf{R}$, there exists a deleted interval about a on which $f > R$. This condition is written $\lim(f, a) = +\infty$ or $\lim\limits_{x \to a} f(x) = +\infty$.

It is a peculiarity of language that if $\lim(f, a) = +\infty$, then $\lim(f, a)$ does not exist; existence of a limit means the limit is a *real number*. Of course, it is possible for a limit not to exist without the limit being $+\infty$; think of the signum function at 0, or the characteristic function of $\mathbf{Q}$ in $\mathbf{R}$.

[1] This is somewhat of an oxymoron, but is standard usage. "Large" refers to "large absolute value."

If f can be made arbitrarily large *negative* by restricting to some deleted interval about a, then we say $\lim(f,a) = -\infty$ or $\lim_{x\to a} f(x) = -\infty$. The precise condition is—aside from one small change—identical to that in Definition 4.35: The conclusion is "$f < R$" instead of "$f > R$."

Certain arithmetic operations involving $+\infty$ can be defined in terms of limits. An equation like "$(+\infty) + (+\infty) = +\infty$" is an abbreviation of, "If f and g have limit $+\infty$ at a, then $f + g$ has limit $+\infty$ at a." In this sense, the following are true (ℓ denotes an arbitrary *positive* real number):

$$\begin{aligned} +\infty \pm \ell &= (+\infty) + (+\infty) = (+\infty)\cdot\ell \\ &= (+\infty)\cdot(+\infty) = +\infty, \qquad \frac{\pm\ell}{(+\infty)} = 0. \end{aligned} \tag{4.6}$$

The proofs are nearly immediate from the definition and are left to you (see Exercise 4.16). The following expressions are *indeterminate*[2] in the sense that the "answer," if it exists at all, depends on the functions f and g, not merely on their limits:

$$(+\infty) - (+\infty), \quad 0\cdot(+\infty), \quad \frac{\ell}{0}, \quad \frac{0}{0}, \quad \frac{(+\infty)}{(+\infty)}.$$

Here are some typical counterexamples, with $a = 0$:

- $(+\infty)-(+\infty)$ $f(x) = 1/|x|$ and $g(x) = 1/|x| - \ell$ for ℓ a fixed real number; each function has limit $+\infty$ at 0, but their difference has limit ℓ. If instead $f(x) = 1/x^2$, then the difference has limit $+\infty$. If $\lim(f,0) = +\infty$ and (say) $g = f - \chi_{\mathbf{Q}}$, then $\lim(g,0) = +\infty$ but $\lim(f-g,0)$ does not exist.
- $0\cdot(+\infty)$ $f(x) = \ell x^2$, $g(x) = 1/x^2$; their product has limit ℓ. If instead $f(x) = \pm|x|$, the product has limit $\pm\infty$. If $f(x) = x$, then the product has no limit at 0.

Projective Infinite Limits

Limits such as $\lim_{x\to 0} 1/x$ do not exist, even allowing $\pm\infty$ as possible values of the limit; in a sense, $1/x$ approaches $-\infty$ as $x \to 0$ from below, but $1/x \to +\infty$ as $x \to 0$ from above. This sort of thing happens whenever $f = p/q$ is a rational function for which q has a root

[2] Or "not consistently definable."

of odd order at a (assuming as usual that p and q have no common factors). One way around this annoyance is to append only a single ∞ to $\mathbf{R}$. This amounts to "gluing" $-\infty$ to $+\infty$ in the extended reals, and is preferable when dealing with rational functions. A deleted interval about ∞ is the *complement* in $\mathbf{R}$ of a closed, bounded interval $[\alpha, \beta]$, and the set $\mathbf{R} \cup \{\infty\}$ together with this notion of "interval about ∞" is the *real projective line*, denoted $\widehat{\mathbf{R}}$ or $\mathbf{RP}^1$. In this context, a function f is said to have limit ∞ at a if, for every $R > 0$, there exists a deleted interval about a on which $|f| > R$. The difference between this and Definition 4.35 is that here we do not care if f is positive or negative, only that it has large absolute value.

There are natural geometric situations in which a single "point at ∞" is better than two. Consider lines through the origin described by their slopes. A line of large positive slope is nearly vertical, but the same is true of a line with large negative slope. It makes sense to interpret lines of slope ∞ as vertical lines, and not to distinguish $+\infty$ and $-\infty$. If we identify the real number m with the line of slope m through $(0, 0)$, then every non-vertical line corresponds to a unique real number, and ∞ corresponds to the vertical axis, see Figure 4.5.

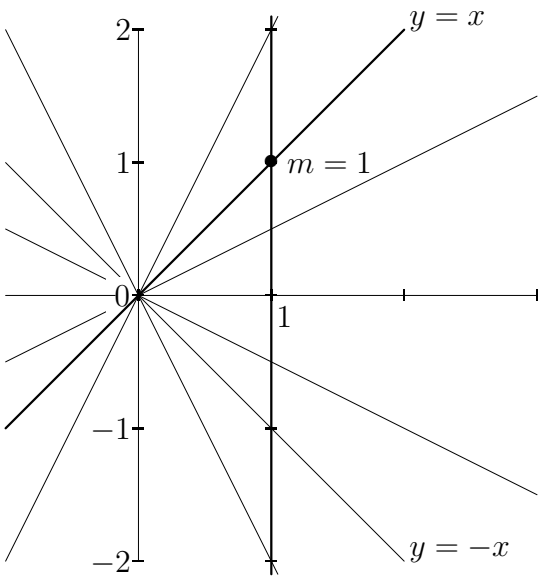


Figure 4.5: Lines through the origin in the plane.

The allowable arithmetic operations with ∞ are different from the allowable operations with $\pm\infty$. For example, the expression $\infty + \infty$

is indeterminate since "$+\infty = -\infty$." The advantage gained is that most divisions by 0 are unambiguous. Precisely, suppose f and g are *rational* functions (g not identically zero) and that $\lim(f, a) = \ell \neq 0$ and $\lim(g, a) = 0$. Then $\lim(f/g, a) = \infty$; briefly, $\ell/0 = \infty$ for $\ell \neq 0$. With similar hypotheses, $\ell/\infty = 0$ and $\ell + \infty = \infty$ for all $\ell \in \mathbf{R}$. The expressions $0/0$, $0 \cdot \infty$, and ∞/∞ are indeterminate; they can be assigned arbitrary value by appropriate choice of f and g.

The proof of Theorem 4.34 shows that if f is a rational function, then $\lim(f, -\infty)$ and $\lim(f, +\infty)$ exist *and are equal*. A very satisfactory picture arises by viewing the domain of f as $\widehat{\mathbf{R}}$ rather than a subset of $\mathbf{R}$: The value of f is ∞ when the denominator is zero, and the value at ∞ is the limiting value (which may itself be ∞). In short, a rational function can be viewed as a mapping $f : \widehat{\mathbf{R}} \to \widehat{\mathbf{R}}$, see Exercise 4.18.

4.3 Continuity

Suppose f is a function defined on an interval about a. There are two numbers—the function value, $f(a)$, and the limit, $\lim(f, a)$—that respectively describe the behavior of f *at* a and *near* a (the limit does not always exist, because the behavior of f near a may be "too complicated" to describe with a single number). For rational functions, we have seen that these two numbers agree.[3] This happy coincidence warrants a name: A function whose domain is an open interval is "continuous at a" if $\lim(f, a) = f(a)$, and is simply called "continuous" if it is continuous at each point of its domain. To define continuity at an endpoint of a closed interval requires one-sided limits. Rather than give a slew of special definitions, we formulate the criterion a bit differently, in a way that makes sense for all functions, regardless of their domain.

Definition 4.36 Let $A \subset \mathbf{R}$ be non-empty. A function $f : A \to \mathbf{R}$ is said to be *continuous at* $a \in A$ if the following condition holds:

> For every $\varepsilon > 0$, there exists a $\delta > 0$ such that if x is a point of A with $|x - a| < \delta$, then $|f(x) - f(a)| < \varepsilon$.

Otherwise f is *discontinuous at* a. If f is continuous at a for every $a \in A$, then we say f is *continuous* (on A).

It is no longer necessary to stipulate $x \neq a$, since if $x = a$ the conclusion $|f(x) - f(a)| < \varepsilon$ is automatic. Let us compare this with

[3] This is even true when the value is ∞, and at the point ∞.

the ordinary limit game; the major changes in the rules are italicized. The referee specifies the function f and the point a, *which must be a point of the domain of* f. The putative limit is taken to be $\ell = f(a)$. Player ε chooses a positive tolerance, and Player δ tries to control the shooting radius so that every shot hits the target. However, Player δ is only required to shoot from points in the domain of f, and *the domain of* f *need not contain a deleted interval about* a.

Consider a sequence $f : \mathbf{N} \to \mathbf{R}$; if $n \in \mathbf{N}$, then the deleted interval of radius 1 about n contains no natural number. Consequently Player δ wins by default because the domain of f contains n but contains no "nearby" points. A sequence is therefore automatically continuous at n for every $n \in \mathbf{N}$. It is clear that if the domain of f does contain an interval about a, then f is continuous at a iff $\lim(f, a) = f(a)$. The reason for making the more general Definition 4.36 is that theorems are easier to state with this definition, and the "true" nature of continuity is not obscured by legalistic questions regarding the domain of f.

Example 4.37 By Corollary 4.21, a rational function is continuous on its natural domain. The signum function is continuous at every point of $\mathbf{R}^\times$, but is discontinuous at 0. The characteristic function of $\mathbf{Q}$ is discontinuous everywhere. The denominator function of Example 3.11 is continuous at every non-rational point and discontinuous at every rational point, by Proposition 4.28. □

Some basic properties of continuous functions follow almost immediately from the analogous results about limits. An obvious modification of the proof of Theorem 4.19 implies that a sum or product of continuous functions is continuous, as is a quotient provided the denominator is non-zero at a. Compositions of continuous functions are almost obviously continuous:

Proposition 4.38. *Let f and g be composable functions. Assume that f is continuous at a and that g is continuous at $f(a)$. Then the composite function $g \circ f$ is continuous at a.*

Proof. Fix $\varepsilon > 0$, and choose $\eta > 0$ such that if y is a point of the domain of g with $|y - f(a)| < \eta$, then $\big|g(y) - g\big(f(a)\big)\big| < \varepsilon$. Then choose $\delta > 0$ such that if x is a point of the domain of f with $|x - a| < \delta$, then $|f(x) - f(a)| < \eta$. This procedure of choosing δ is a winning strategy in the continuity game for $g \circ f$ at a, since if x is in the domain of $g \circ f$ (i.e., the domain of f), then

$$|x - a| < \delta \implies |f(x) - f(a)| < \eta \implies |(g \circ f)(x) - (g \circ f)(a)| < \varepsilon;$$

thus $g \circ f$ is continuous at a. □

Let f be a function whose domain contains an interval about a. Possible discontinuities of f at a are often categorized into three types: The point a is

- A *removable discontinuity* if $\lim(f, a)$ exists but is not equal to $f(a)$. In this case, "$f(a)$ has the wrong value." By redefining f at a to be $\lim(f, a)$, the discontinuity is removed. It is not uncommon to say that a point where $\lim(f, a)$ exists is a removable discontinuity, even if f is undefined at a. This is usually harmless, but not strictly correct.

- A *jump discontinuity* if the one-sided limits exist but are not equal to each other (one of them may be $f(a)$, or not). Intuitively, the graph of f "jumps" from $\lim(f, a^-)$ to $\lim(f, a^+)$ at a, see Figure 4.3.

- A *wild*[4] *discontinuity* if at least one of the one-sided limits fails to exist.

The function $\chi_{\{0\}}$ (the characteristic function of the singleton $\{0\}$) has a removable discontinuity at 0, the signum function has a jump discontinuity at 0, and the reciprocal function $x \mapsto 1/x$ has a wild discontinuity at 0. A function may have infinitely many discontinuities of each type in a bounded interval. The denominator function has a removable discontinuity at every rational number. A non-decreasing function has only jump discontinuities by Theorem 4.30, and it is not difficult to arrange that there are infinitely many of them. Finally, $\chi_{\mathbf{Q}}$ has a wild discontinuity at every real number. There are subtle restrictions on the discontinuity set that are beyond the scope of this book. For example, it is impossible that the discontinuity set of $f : \mathbf{R} \to \mathbf{R}$ is exactly the set of irrational numbers.

Continuity is a local property, that is, it depends only on the behavior of a function near a point. Many of the most interesting results about continuous functions are *global*; they depend on the behavior of the function everywhere in its domain. Some of these are introduced in Chapter 5.

[4]This is not a standard term.

4.4 Sequences and Series

Sequences and their limits are one of the most important topics in analysis, both theoretically and in applications. Arguably, convergence of a sequence is the simplest way a function can have a limit. At the same time, sequences arise in interesting ways, such as iteration of maps, continued fractions, and infinite sums.

Let $(a_n)_{n=0}^{\infty}$ be a sequence. According to the definition, the sequence has a *limit* $\ell \in \mathbf{R}$ if, for every $\varepsilon > 0$, there is an $N \in \mathbf{N}$ such that $|a_n - \ell| < \varepsilon$ for $n > N$. This condition has a simple formulation with no analogue for functions on intervals:

> A sequence $(a_n)_{n=0}^{\infty}$ converges to ℓ iff every open interval about ℓ contains all but finitely many of the terms a_n.

The terms are counted according to the number of times they appear in the infinite list a_0, a_1, a_2,..., not according to the number of points in the image. For example, the sequence defined by $a_n = (-1)^n$ has image $\{-1, 1\}$, and every open interval about 1 contains all but finitely many points of the image. However, an open interval of radius smaller than 2 fails to contain all the odd terms a_{2k+1}, $k \in \mathbf{N}$, and there are infinitely many of these. Consequently this sequence does not converge to 1 (nor to any other ℓ).

If (a_n) is a sequence in A and $f : A \to \mathbf{R}$ is a function, then the composition of f with a is a real sequence (b_n), with $b_n = f(a_n)$. Convergence of such sequences can be used to determine whether or not f has a limit; for technical reasons, one uses sequences that do not hit the point a.

Theorem 4.39. *Let $f : A \to \mathbf{R}$ be a function whose domain contains a deleted interval about a. Then $\lim(f, a)$ exists and is equal to $\ell \in \mathbf{R}$ iff $\lim_{n\to\infty} f(a_n) = \ell$ for every sequence (a_n) in $A \setminus \{a\}$ that converges to a.*

Proof. Suppose $\lim(f, a) = \ell$, and let (a_n) be a sequence in $A \setminus \{a\}$ that converges to a. Fix $\varepsilon > 0$ and choose $\delta > 0$ so that if $0 < |x - a| < \delta$, then $|f(x) - \ell| < \varepsilon$. Then choose $N \in \mathbf{N}$ such that if $n > N$, then $0 < |a_n - a| < \delta$; this is possible because the sequence (a_n) converges to a but is never equal to a. If $n > N$, then these inequalities imply $|f(a_n) - \ell| < \varepsilon$, so that $f(a_n) \to \ell$ as $n \to \infty$. (Compare with the proof of Proposition 4.38.)

Conversely, suppose "$\lim(f, a) = \ell$" is false, that is the limit exists but is not equal to ℓ or the limit does not exist. In the game-theoretic

interpretation, Player ε has a winning strategy. Let us follow the course of a game using the interpretation in which Player ε chooses some $\varepsilon > 0$, Player δ chooses a $\delta > 0$, and finally Player ε chooses a point x with $0 < |x - a| < \delta$. First Player ε chooses a sufficiently small $\varepsilon > 0$. Because Player δ cannot win against the choice of ε, there exists, for each natural number n, a point $a_n \in A \setminus \{a\}$ with

$$|a_n - a| < \frac{1}{n} \quad \text{and} \quad |f(a_n) - \ell| \geq \varepsilon;$$

the point a_n is a winning choice for Player ε against $\delta = 1/n$. Now consider the sequence $(a_n)_{n=1}^{\infty}$. By construction none of the terms is equal to a, but since $|a_n - a| < 1/n$ the sequence converges to a. Finally, there is an $\varepsilon > 0$ such that $|f(a_n) - \ell| \geq \varepsilon$ for all $n \in \mathbf{N}$, so the sequence (b_n) with $b_n = f(a_n)$ *does not* converge to ℓ. □

An obvious modification of the proof works for limits at $+\infty$ or $-\infty$. Theorem 4.39 is useful for proving non-existence of one-sided limits, a task that can otherwise be messy. For example, let $f : \mathbf{R} \to \mathbf{R}$ be a non-constant periodic function (such as the "Charlie Brown" function of Example 3.25), and define $g : \mathbf{R}^{\times} \to \mathbf{R}$ by $g(x) = f(1/x)$. Then $\lim(g, 0)$ does not exist. Intuitively, the function g oscillates infinitely many times on each interval $(0, \delta)$, because the function f oscillates infinitely many times on the interval $(1/\delta, +\infty)$. To give a formal proof, construct two sequences of positive numbers, say (a_n) and (b_n), that converge to 0 but so that the corresponding sequences of function values have different limits. For definiteness, say the period of f is α. Choose points x and y in $(0, \alpha]$ such that $f(x) \neq f(y)$; by periodicity,

$$f(x + n\alpha) = f(x) \neq f(y) = f(y + n\alpha) \qquad \text{for all } n \in \mathbf{N}.$$

Set $a_n = 1/(x + n\alpha)$ and $b_n = 1/(y + n\alpha)$; these sequences have the desired properties, as you should verify.

The proof of Theorem 4.39 is easily modified to establish the following useful consequence. Conceptually, evaluating a continuous function commutes with taking the limit of a sequence, see equation (4.7). This property will be used repeatedly in applications.

Corollary 4.40. *Let $f : A \to \mathbf{R}$ be a function. Then f is continuous at $a \in A$ iff*

$$\lim_{n\to\infty} f(a_n) = f\Big(\lim_{n\to\infty} a_n\Big) \tag{4.7}$$

for every sequence $(a_n)_{n=1}^{\infty}$ in A that converges to a.

Theorem 4.41 is an analogue of Theorem 4.30 for sequences; the proof is left as an exercise. Such a theorem is useful for proving convergence of a sequence when the limit cannot be guessed in advance. Of course, a similar result holds for non-increasing sequences.

Theorem 4.41. *Let (a_n) be a non-decreasing sequence of real numbers. Then (a_n) converges iff it is bounded above.*

Cauchy Sequences

As has been hinted already, convergence of a sequence can be an unwieldy theoretical condition, because it cannot be verified for a specific sequence unless the limit is known, or unless some other information is available (such as "the sequence is monotone and bounded"). It would be useful to have a general convergence criterion that does not require knowledge of the limit. The "Cauchy[5] criterion" fulfills this purpose.

Definition 4.42 A sequence $(a_n)_{n=0}^{\infty}$ is a *Cauchy sequence* if, for every $\varepsilon > 0$, there exists an $N \in \mathbf{N}$ such that $m, n \geq N$ implies $|a_n - a_m| < \varepsilon$.

Intuitively, the terms of a Cauchy sequence can be made arbitrarily close *to each other* by going sufficiently far out in the sequence. This condition does not even depend on existence of a limit, much less on the exact *value* of the limit. By contrast, convergence of a sequence means the terms can be made arbitrarily close *to a fixed number* (the limit) by going sufficiently far out in the sequence. The difference is subtle, but important. Before reading further, consider briefly whether every Cauchy sequence is convergent, or whether every convergent sequence is Cauchy, or both, or neither. As a hint, you should have no trouble resolving one direction.

To give you a feel for the Cauchy criterion, here two basic applications.

Lemma 4.43. *If (a_n) is a Cauchy sequence, then there exists $R \in \mathbf{R}$ such that $|a_n| \leq R$ for all $n \in \mathbf{N}$. Briefly, a Cauchy sequence is bounded.*

Proof. Let $\varepsilon = 1$. Because (a_n) is Cauchy, there exists $N \in \mathbf{N}$ such that $|a_n - a_m| < 1$ for $n, m \geq N$. Setting $m = N$ and using the triangle

[5] KOH shee; in French, the *second* syllable is accented.

inequality, $|a_n| = |(a_n - a_N) + a_N| \leq 1 + |a_N|$ for all $n \geq N$. The desired conclusion follows by taking

$$R = \max(|a_1|, |a_2|, \ldots, |a_{N-1}|, 1 + |a_N|),$$

the maximum of a finite list of numbers. □

Lemma 4.44. *If (a_n) converges, then (a_n) is Cauchy.*

Proof. Let ℓ denote the limit of (a_n). Fix $\varepsilon > 0$, and choose $N \in \mathbf{N}$ such that $|a_n - \ell| < \varepsilon/2$ for $n \geq N$. By the triangle inequality,

$$|a_n - a_m| \leq |a_n - \ell| + |a_m - \ell| < \varepsilon$$

for m, $n \geq N$. □

The "converse" question "Does every Cauchy sequence converge?" is more subtle; at issue is the construction of a putative limit ℓ from a Cauchy sequence. The completeness property turns out to be crucial.

Theorem 4.45. *Let $(a_k)_{k=0}^{\infty}$ be a Cauchy sequence of real numbers. There exists $\ell \in \mathbf{R}$ such that $\lim_{k \to \infty} a_k = \ell$.*

Proof. Construct an auxiliary sequence $(b_n)_{n=0}^{\infty}$ as follows:

$$b_n = \sup\{a_k \mid k \geq n\}. \tag{4.8}$$

In words, look at the successive "tails" of (a_k), and take their "maxima". The sequence (b_n) is clearly non-increasing: You cannot make the supremum of a set larger by removing elements! Further, Lemma 4.43 says (a_k), and hence (b_n), is bounded. Theorem 4.41 says that (b_n) has a real limit, which we call ℓ.

Fix $\varepsilon > 0$ and use the Cauchy condition for (a_k) to choose $N_0 \in \mathbf{N}$ such that $|a_n - a_m| < \varepsilon$ for n and $m \geq N_0$. Next, choose $N_1 \geq N_0$ such that $|b_n - \ell| < \varepsilon$ for $n \geq N_1$. Now, $b_{N_1} = \sup\{a_k \mid k \geq N_1\}$, so there exists $N \geq N_1$ such that $|b_{N_1} - a_N| < \varepsilon$. If $n \geq N$, then

$$|a_n - \ell| \leq |a_n - a_N| + |a_N - b_{N_1}| + |b_{N_1} - \ell| < 3\varepsilon.$$

This means $(a_k) \to \ell$. □

Cantor's Construction of R

As a short detour (that is not used elsewhere in the book), here is a sketch of Cantor's construction of the real numbers. First, note that the concept of Cauchy sequences makes sense for sequences of rational numbers, even if one does not know about real numbers; the Cauchy criterion is defined purely in terms of the sequence itself.

A salient deficiency of $\mathbf{Q}$ is that there exist Cauchy sequences in $\mathbf{Q}$ that do not converge—that do not have a *rational* limit. The sequence of Example 4.47 below with $x_0 = b = 2$ (cf. Lemma 2.29) is a Cauchy sequence of rational numbers, but has no rational limit, as there does not exist a rational number whose square is 2. If one lives in $\mathbf{Q}$, one can only agree that such a sequence diverges.

Naively, one would like to *define* a real number to be a Cauchy sequence of rational numbers; this makes sense purely in terms of $\mathbf{Q}$, and with the hindsight that a Cauchy sequence of real numbers converges, we would identify a sequence with its (real) limit. The hitch is that many Cauchy sequences have the same limit, so a real number should be identified with the *set* of Cauchy sequences that have the same real limit. Unfortunately, this "definition" no longer refers to $\mathbf{Q}$ alone, so we must reformulate it.

Cantor declared two Cauchy sequences, (a_n) and (b_n), to be *equivalent* if $\lim_{n\to\infty} |a_n - b_n| = 0$. It is easy to see that this is an equivalence relation on the set of Cauchy sequences of rationals. (Reflexivity and symmetry are obvious, and transitivity follows from the triangle inequality.) As you might guess, Cantor defines a *real number* to be an *equivalence class of Cauchy sequences* under this equivalence relation. The beauty of this definition is that the field axioms and order properties can be checked using theorems we have already proven (namely, that limits preserve arithmetic and order relations, see Theorems 4.19 and 4.23).

Completeness is the only remaining property, and it is verified by a "diagonal argument": If $A_m := (a_{n,m})_{n=0}^{\infty}$ is a Cauchy sequence for each $m \in \mathbf{N}$ (that is, $(A_m)_{m=0}^{\infty}$ is a sequence of Cauchy sequences!), and if (A_m) is non-decreasing and bounded above in the sense of the order properties, then (one argues) there exists an increasing set of indices $(n_k)_{k=0}^{\infty}$ such that the rule $b_k = a_{k,n_k}$ defines a Cauchy sequence that corresponds to the supremum of the set $\{A_m\}$. Modulo details, this proves completeness.

If you compare this construction with the procedures in Chapter 2

by which the integers, rational, real, and complex numbers were successively constructed, you will find that the construction of the real numbers from the rationals is the most complicated step, involving infinite sets (cuts) of rationals or equivalence classes of Cauchy sequences. You might wonder whether there is a simpler construction, involving only finite sets of rationals, or perhaps equivalence classes of finite sets. (Each of the other constructions uses equivalence classes of *pairs* of numbers of the previous type.) The answer is provably *no*. With a small modification, Theorem 3.19 (due to Cantor) states:

> Let $\mathbf{Q}^n$ be the set of ordered n-tuples of rational numbers (a.k.a., rational sequences of length n), and let $\mathbf{Q}^\infty$ denote the union of the $\mathbf{Q}^n$ over $n \in \mathbf{N}$ (a.k.a., the set of all finite sequences of rational numbers). There does not exist a surjective function $f : \mathbf{Q}^\infty \to \mathbf{R}$.

It is consequently impossible to construct the real numbers from the rational numbers using only finite sets of rationals; *there are not enough finite sets of rationals*!

Limits of Recursive Sequences

Let $X \subset \mathbf{R}$ be an interval (for convenience), and let $f : X \to X$ be a *continuous* function. The iterates of f determine a *discrete dynamical system* on X, and the *orbit* of a point $x_0 \in X$ is the sequence $(x_n)_{n=0}^\infty$ defined by $x_{n+1} = f(x_n)$ for $n \geq 0$. There is a geometric way to visualize the orbit of a point, see Figure 4.6. Draw the graph of f in the square $X \times X$, and draw the diagonal. Start with x_0 on the horizontal axis. Go up to the graph of f, then left or right to the diagonal. Repeat, going up or down to the graph of f and left or right to the diagonal. The horizontal positions of the points generated are the iterates of x_0.

The present question of interest is, "If (x_n) converges, what is the limit?" The nice answer is that such a sequence must converge to a *fixed point* of f, namely a point ℓ with $f(\ell) = \ell$. The formalism of limits and continuity can be used to give a simple proof: If $\lim_{n\to\infty} x_n = \ell$, then

$$f(\ell) = f\Big(\lim_{n\to\infty} x_n\Big) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} x_n = \ell.$$

The first, third, and fifth equalities are definitions. The second equality is Corollary 4.40, and the fourth is clear from the definition of limit.

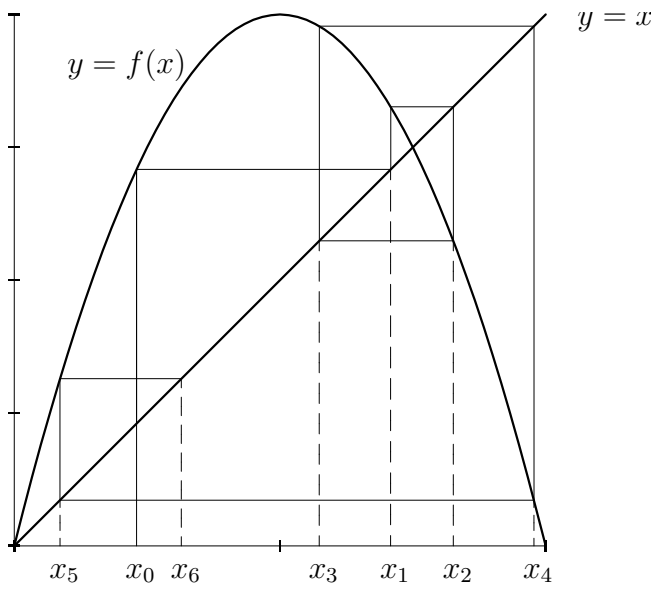


Figure 4.6: The orbit of a point under iteration of f.

While this observation does not always locate the limit of a recursive sequence, it reduces the possible choices to a finite set in many interesting situations.

Example 4.46 Let $f : [0,1] \to [0,1]$ be defined by $f(x) = x^2$. The fixed points of f are 0 and 1, so if the orbit of x_0 under f converges, it must converge to one of these points. If $0 \leq x \leq 1$, then $x^2 \leq x$, so for each $x_0 \in [0,1]$ the orbit of x_0 is a non-increasing sequence, which converges by Theorem 4.41. For this function, it is easy to see that the orbit of 1 is the constant sequence $x_n = 1$ for all n, while for $x_0 < 1$ the orbit of x_0 converges to 0.

For the function $g : [-1,1] \to [-1,1]$ defined by $g(x) = -x$, the only fixed point is 0. If $x_0 \neq 0$, then the orbit of x_0 does not converge; in this example, knowledge of the fixed points is almost useless. □

Example 4.47 Let $b > 0$; we will prove that b has a real square root, namely that there exists a positive real number ℓ with $\ell^2 = b$. Details and further information are relegated to Exercise 4.23.

Fix $x_0 > b$, and define a sequence recursively by

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{b}{x_n}\right) \qquad \text{for } n \in \mathbf{N}. \tag{4.9}$$

Algebra and induction on n imply that (x_n) is bounded below by 0 and decreasing. By Theorem 4.41, (x_n) converges to a real number $\ell \neq 0$. Because $x_n > 0$ for all n, Theorem 4.23 implies that $\ell \geq 0$, hence $\ell > 0$. On the other hand, the sequence is gotten by iterating the continuous function $f : (0, \infty) \to (0, \infty)$ defined by

$$f(x) = \frac{1}{2}\left(x + \frac{b}{x}\right),$$

so ℓ is a fixed point of f. The fixed point equation $f(\ell) = \ell$ is equivalent to $\ell^2 = b$, which proves that b has a square root, henceforth denoted $\sqrt{b}$. As it turns out, this sequence converges very rapidly to $\sqrt{b}$; Exercise 4.23 provides an estimate on the rate of convergence. □

This example typifies the "solution" of a problem in analysis. Some kind of infinite object (in this case a sequence) is considered, and some piece of information (in this case the limit) is wanted. Existence of the limit is proven by a general result; the hypotheses of the relevant theorem(s) must be verified for the particular case at hand. A separate result (the fixed-point result, in this case) narrows down the limit to a finite number of possibilities, and some additional work dispenses with all but one choice.

If you have taken a calculus course already, you are familiar with at least parts of this outline. To find the maximum value of a "differentiable" function, you set the derivative equal to 0 to rule out all but a (usually) finite number of possible locations for the maximum. Some additional work eliminates all the wrong choices. The remaining piece is the *existence* of a maximum value; an appropriate existence result is proven in Chapter 5.

Infinite Series

Every sequence $(a_k)_{k=0}^{\infty}$ (of real or complex numbers) is associated to a sequence $(s_n)_{n=0}^{\infty}$ of *partial sums* (the "running total"), defined by

$$s_n = \sum_{k=0}^{n} a_k = a_0 + a_1 + \cdots + a_n. \tag{4.10}$$

The original sequence can be recovered from the partial sums via

$$a_0 = s_0, \qquad a_n = s_n - s_{n-1} \quad \text{for } n \geq 1,$$

so you might wonder why we bother with sums at all. The reason is simply that many mathematical problems present themselves naturally as sums rather than as terms in a sequence. The identity

$$s_n - s_m = \sum_{k=m+1}^{n} a_k, \qquad 0 \leq m < n,$$

is useful and will be used repeatedly.

Definition 4.48 Let $(a_k)_{k=0}^{\infty}$ be a sequence of real numbers. If the sequence $(s_n)_{n=0}^{\infty}$ of partial sums is convergent, then (a_k) is said to be *summable*. The limit is denoted

$$\sum_{k=0}^{\infty} a_k := \lim_{n\to\infty} s_n, \tag{4.11}$$

and is called the *sum of the series*.

Alternatively, the "series" $\sum_k a_k$ is said to "converge." This usage is convenient because it is common to write an expression $\sum_k a_k$ even if the sequence (a_k) is not known to be summable. It is crucial to remember, though, that convergence of a sequence (a_k) is quite different from convergence of the series $\sum_k a_k$; the only general relationship between the two concepts is given in Proposition 4.51.

Series behave as expected with respect to termwise addition, and scalar multiplication. Products of series are more subtle, and are treated only under an additional hypothesis. You should have no trouble proving the following from the definitions.

Theorem 4.49. *If (a_k) and (b_k) are summable and $c \in \mathbf{R}$, then the sequences $(a_k + b_k)$ and (ca_k) are summable, and*

$$\sum_{k=0}^{\infty}(a_k + b_k) = \sum_{k=0}^{\infty} a_k + \sum_{k=0}^{\infty} b_k, \qquad \sum_{k=0}^{\infty}(ca_k) = c\sum_{k=0}^{\infty} a_k.$$

Summing a sequence may be regarded as "accounting with an infinite ledger." You have an infinite list of numbers a_0, a_1, a_2, ..., add them up in succession, and ask whether the running totals approach a limit. If so, the sequence is summable and the limit is (loosely) viewed as the "sum" of the numbers in the list, *taken in the order presented.* (This proviso is important.) Summability of a sequence depends only on the tail of the sequence. Said another way, discarding finitely many

terms of a sequence does not change summability (though of course it *does* generally change the actual sum). Clearly, terms that vanish can be deleted without affecting either summability or the actual sum. Further, arbitrarily many zero terms can be shuffled into the list, so long as all the original terms remain in the same order.

Summability is a subtle and slightly counterintuitive property. In the infinite ledger, delete all terms that vanish, and divide the terms into credits (positive terms) and debits (negative terms). If the original sequence is summable, then we can deduce one of two things:

- The credits and debits are separately summable (have finite sum), or
- The credits and debits separately fail to be summable (have infinite sum).

In the former case, the sum is insensitive to rearrangement (this takes some work to prove), but in the latter case the sum of the ledger is something like the indeterminate expression $(+\infty) - (+\infty)$, and the order of the terms is important; "rearrangement" of the terms of a sequence *can alter summability, and can change the value of the sum*! Despite the notation, a series is not really a sum of infinitely many terms, but a *limit of partial sums*.

We have already encountered two instances of summable sequences: Geometric series (Exercise 2.16) and decimal representations of real numbers. For the record, here is the full story on geometric series.

Proposition 4.50. *Let $a \neq 0$, $r \in \mathbf{R}$. The sequence $(ar^k)_{k=N}^{\infty}$ is summable iff $|r| < 1$, and in this event*

$$\sum_{k=N}^{\infty} ar^k = \frac{ar^N}{1-r}.$$

Proof. For each $n \in \mathbf{N}$,

$$\sum_{k=N}^{N+n} ar^k = ar^N \sum_{j=0}^{n} r^j.$$

Recall from Example 2.20 that

$$(*) \qquad \sum_{j=0}^{n} r^j = \begin{cases} \dfrac{1-r^{n+1}}{1-r} \text{ if } r \neq 1, \\ \\ (n+1) \text{ if } r = 1. \end{cases}$$

If $|r| > 1$ or $r = 1$, then the partial sums grow without bound (in absolute value), so the series does not converge. When $r = -1$, the partial sums are alternately 1 and 0, so the series does not converge in this case.

Finally, if $|r| < 1$, then $r^n \to 0$ as $n \to \infty$ by Exercise 2.5 (c), so the series in equation $(*)$ has sum $1/(1-r)$. The desired conclusion follows by Theorem 4.49. □

Geometric series are among the most elementary series, because they can be summed explicitly; that is, the series is not merely known to converge, but the sum can be calculated. Decimal representations are special because their terms are non-negative, so the sequence of partial sums is non-decreasing. In general, convergent series are not so well-behaved (neither explicitly summable nor having monotone partial sums), so it is desirable to have general theoretical tools for determining convergence. The basic tool in deriving convergence tests for series is the Cauchy criterion, applied to the sequence of partial sums. The very simplest example of a convergence condition is the *vanishing criterion*:

Proposition 4.51. *If $(a_k)_{k=0}^{\infty}$ is a summable sequence, then $\lim_{k\to\infty} a_k = 0$.*

Proof. Summability of (a_k) means, by definition, that (s_n), the sequence of partial sums, is convergent, hence Cauchy by Lemma 4.44. Taking $m = n+1$ in the definition of "Cauchy sequence" says that for all $\varepsilon > 0$, there is an $N \in \mathbf{N}$ such that $n \geq N$ implies $|a_{n+1}| = |s_m - s_n| < \varepsilon$. This means exactly that $\lim_{n\to\infty} a_n = 0$. □

It cannot be emphasized too heavily that there is no "universal" test for convergence or divergence that works for all series, nor is there a procedure for evaluating the sum of a sequence that is known to be summable. In particular, the converse of Proposition 4.51 is *false*: It is possible for a sequence to converge to 0 without being summable. For sequences of positive terms, summability measures—in a subtle way—the "rate" at which the terms go to 0.

Comparison Tests

Almost all tests for convergence of a series rely on comparison with a series known to converge. The next theorem is the *comparison test*, and Corollary 4.54 is the *limit comparison test*.

Theorem 4.52. *Let (b_k) be a summable sequence of* non-negative *terms. If (a_k) is a sequence with $|a_k| \leq b_k$ for all $k \in \mathbf{N}$, then (a_k) is summable.*

Because summability is a property of the tail of (a_k), the hypothesis can be weakened to "There exists $N \in \mathbf{N}$ such that $|a_k| \leq b_k$ for $k \geq N$." The contrapositive is a *divergence* test: If $(|a_k|)$ is *not* summable, then (b_k) is also not summable.

Proof. Let (s_n) be the sequence of partial sums of (a_k), and let (t_n) be the sequence of partial sums of (b_k). By the triangle inequality, if $m < n$ then

$$\begin{aligned} |s_n - s_m| = \left| \sum_{k=m+1}^{n} a_k \right| &\leq \sum_{k=m+1}^{n} |a_k| \\ &\leq \sum_{k=m+1}^{n} b_k = |t_n - t_m|. \end{aligned}$$

Because (b_k) is summable, (t_n) is Cauchy, so the previous estimate shows (s_n) is also Cauchy, i.e., (a_k) is summable. □

Theorem 4.53. *Let (b_k) be a summable sequence of* positive *terms. If (a_k) is a sequence such that $\lim_{k\to\infty}(a_k/b_k)$ exists, then (a_k) is summable.*

Proof. If $a_k/b_k \to \ell$ as $n \to \infty$, then $|a_k|/b_k \to |\ell|$ as $n \to \infty$. To see this, apply Corollary 4.40 to the absolute value function. By Theorem 4.52, it suffices to show $(|a_k|)$ is summable. Pick a real number $M > |\ell|$ and write $M = |\ell| + \varepsilon$ with $\varepsilon > 0$. Choose $N \in \mathbf{N}$ such that

$$n \geq N \implies \left| \frac{|a_k|}{b_k} - |\ell| \right| < \varepsilon.$$

A bit of algebra shows that if $n \geq N$, then $|a_k| \leq Mb_k$. The sequence (Mb_k) is summable by Theorem 4.49, so (a_k) is summable by Theorem 4.52. □

Corollary 4.54. *If (a_k) and (b_k) are sequences of* positive *terms, and if*

$$\lim_{k\to\infty} \frac{a_k}{b_k} = \ell \neq 0,$$

then (a_k) is summable iff (b_k) is summable.

Absolute Summability

As suggested earlier, summability involves two technical issues: Do the positive terms and negative terms *separately* have finite sum, or is there fortuitous cancellation based on the ordering of the summands? In this section we study the former case.

Definition 4.55 Let (a_k) be a summable sequence. If $(|a_k|)$ is also summable, then we say that (a_k) is *absolutely summable*, and that the series $\sum_k a_k$ is *absolutely convergent*. If $(|a_k|)$ is not summable, then (a_k) is *conditionally summable*, and $\sum_k a_k$ is *conditionally convergent*.

Every sequence (a_k) falls into exactly one of the following categories: not summable, conditionally summable, or absolutely summable. Clearly, a sequence of non-negative terms is either absolutely summable or not summable.

Given a real sequence (a_k), define the associated sequences of positive and negative terms by

$$a_k^+ = \max(a_k, 0), \qquad a_k^- = \min(a_k, 0). \tag{4.12}$$

For example, if $a_k = (-1)^k k$, then

k	0	1	2	3	4	5	6	$\cdots$
a_k	0	-1	2	-3	4	-5	6	$\cdots$
a_k^+	0	0	2	0	4	0	6	$\cdots$
a_k^-	0	-1	0	-3	0	-5	0	$\cdots$

Proposition 4.56. *A real sequence (a_k) is absolutely summable iff both sequences $(a_k^\pm)$ are summable.*

Proof. First observe that $|a_k^+| = a_k^+$ and $|a_k^-| = -a_k^-$ for all k; consequently, the sequences of positive or negative terms are summable iff they are absolutely summable.

Now, $|a_k^\pm| \leq |a_k|$ for all k, so by Theorem 4.52 (the comparison test) if (a_k) is absolutely summable, then so are $(a_k^\pm)$. Conversely, $|a_k| = a_k^+ - a_k^-$ for all k, so if $(a_k^\pm)$ are summable, then (a_k) is absolutely summable. □

Let $m : \mathbf{N} \to \mathbf{N}$ be a bijection, and let (a_k) be a sequence. The sequence defined by $b_k = a_{m(k)}$ is a *rearrangement*. Informally, a rearrangement is the same ledger sheet, read in a different order. Note,

however, that rearranging cannot do things like "read all the even terms, then all the odd terms", since the list of even terms would already be an infinite list. Our next aim is to show that "rearranging an absolutely summable sequence does not alter the sum".

Theorem 4.57. *Let (a_k) be absolutely summable, and let (b_k) be a rearrangement. Then (b_k) is absolutely summable, and $\sum_k b_k = \sum_k a_k$.*

The idea of the proof is simple: Take enough terms a_k to approximate $\sum_k a_k$ closely, then go far enough out in the rearrangement to include all the chosen terms. The corresponding partial sum of (b_k) is close to $\sum_k a_k$, too.

Proof. Let A_n be the nth partial sum of (a_k), B_n the nth partial sum of (b_k), and $A = \lim_n A_n$. We wish to show that $(B_n) \to A$.

Fix $\varepsilon > 0$, and use absolute summability to choose N_0 so that

$$\sum_{k=0}^{\infty} |a_k| - \sum_{k=0}^{N_0} |a_k| = \sum_{k=N_0+1}^{\infty} |a_k| < \frac{\varepsilon}{2}.$$

In particular, $|A - A_{N_0}| < \varepsilon/2$. Now choose $N \in \mathbf{N}$ such that the summands $a_0, \ldots, a_{N_0}$ are among the terms $b_0, b_1, \ldots, b_N$; this is possible because (b_k) is a rearrangement of (a_k). If $n \geq N$, then

$$|B_n - A_N| = \left| \Big(\sum_{k=0}^{n} b_k \Big) - (a_0 + a_1 + \cdots + a_N) \right| \leq \sum_{k=N+1}^{\infty} |a_k| < \frac{\varepsilon}{2},$$

so $|B_n - A| \leq |B_n - A_N| + |A_N - A| < \varepsilon$. □

Our last general result about absolute summability concerns products of series. Let (a_k) and (b_ℓ) be summable sequences. We want to make sense of the "infinite double sum" $\sum_{k,\ell} a_k b_\ell$, and if possible to evaluate this expression in terms of the sums of (a_k) and (b_ℓ). The first problem is that the sum is taken over $\mathbf{N} \times \mathbf{N}$, and while this set is countable, there is no "natural" enumeration we can use to get a series. The second problem is that the "sum" might turn out to depend on the enumeration we pick, just as a series' value may change under rearrangement, or might fail to exist at all. If neither of (a_k) nor (b_ℓ) is absolutely summable, these potential snags are genuine difficulties. If either sequence *is* absolutely summable, then the double sum is the

product of the individual sums; however, we do not need this generality, and will only treat the case where both sequences are absolutely summable.

The "standard ordering" of the double series is provided by the *Cauchy product* of (a_k) and (b_ℓ), the iterated sum

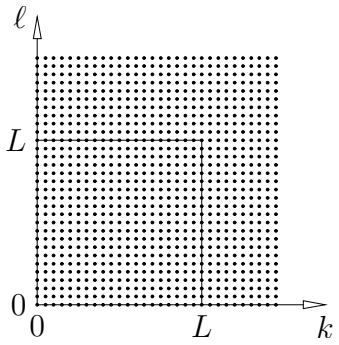
$$\sum_{n=0}^{\infty} \sum_{k+\ell=n} a_k b_\ell = \sum_{n=0}^{\infty} \sum_{k=0}^{n} a_k b_{n-k}. \tag{4.13}$$

The Cauchy product is extremely useful later, when we use it to multiply power series.

Theorem 4.58. *Let (a_k) and (b_ℓ) be absolutely summable sequences of real numbers, and let (c_m) be an arbitrary enumeration of the doubly-infinite set $\{a_k b_\ell \mid (k,\ell) \in \mathbf{N} \times \mathbf{N}\}$. Then (c_m) is absolutely summable, and*

$$\sum_{m=0}^{\infty} c_m = \Big(\sum_{k=0}^{\infty} a_k\Big)\Big(\sum_{\ell=0}^{\infty} b_\ell\Big).$$

The space of indices of the double series may be visualized as follows:



As in Theorem 4.57, the idea is that "most" of the contribution to the sum comes from terms inside the lower left square; some finite initial portion of (c_n) includes these terms, and the sum of the remaining terms is small.

Proof. It greatly simplifies the notation to introduce the following:

$$A_n = \sum_{k=0}^{n} a_k, \qquad B_n = \sum_{\ell=0}^{n} b_\ell, \qquad C_n = \sum_{m=0}^{n} c_m,$$

$$\mathbf{A}_n = \sum_{k=0}^{n} |a_k|, \qquad \mathbf{B}_n = \sum_{\ell=0}^{n} |b_\ell|.$$

Finally, let $A = \lim_n A_n$, $B = \lim_n B_n$, $\mathbf{A} = \lim_n \mathbf{A}_n$, and $\mathbf{B} = \lim_n \mathbf{B}_n$.

The sequence $P_n := A_n B_n$ converges to AB, since the limit of a product is the product of the limits. Similarly, $\mathbf{P}_n = \mathbf{A}_n \mathbf{B}_n$ converges to $\mathbf{AB}$.

Fix $\varepsilon > 0$, and choose $L \in \mathbf{N}$ such that

$$(*) \qquad |AB - P_n| < \varepsilon \quad \text{and} \quad |\mathbf{AB} - \mathbf{P}_n| < \varepsilon \qquad \text{for } n \geq L.$$

Now choose $N \geq L$ so that every term "inside the square", namely every product $a_k b_\ell$ with k, $\ell < L$, is among the terms c_0, c_1, ..., c_N. If $n \geq N$, then

$$|C_n - P_n| = |C_n - A_n B_n| \leq \sum_{k \text{ or } \ell > L} |a_k| \cdot |b_\ell| = |\mathbf{AB} - \mathbf{P}_n| < \varepsilon,$$

by the second part of $(*)$. Consequently, if $n \geq N$, then

$$|AB - C_n| \leq |AB - P_n| + |P_n - C_n| < 2\varepsilon$$

by the first part of $(*)$. □

The Ratio and Root Tests

The tests above presume that a sequence known to be summable is available. The next two tests, the *ratio* and *root* tests, can be used to prove specific sequences are summable, by comparing them to a geometric series. Unfortunately, neither test is widely applicable, though both are useful for "power series," see Chapter 11.

Theorem 4.59. *Let (a_k) be a sequence of* positive *terms, and assume*

$$\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = \rho$$

exists. If $\rho < 1$, then (a_k) is absolutely summable. If $\rho > 1$, then (a_k) is not summable.

Proof. Suppose $\rho < 1$. Choose $r \in (\rho, 1)$, and set $\varepsilon = r - \rho > 0$. By hypothesis, there is an $N \in \mathbf{N}$ such that (see Figure 4.7)

$$n \geq N \implies \left| \frac{a_{n+1}}{a_n} \right| < \rho + \varepsilon = r,$$

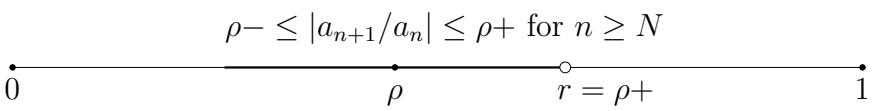


Figure 4.7: Bounding ratios in the Ratio Test

or $|a_{n+1}| \le |a_n|r$ for $n \ge N$. By induction on k,

$$|a_{N+k}| \le |a_N|r^k \qquad \text{for } k \in \mathbf{N}.$$

Consequently, the tail $(a_{N+k})_{k=0}^{\infty}$ is bounded above in absolute value by $(|a_N|r^k)_{k=0}^{\infty}$, a convergent geometric series.

If $\rho > 1$, then choosing $r \in (1, \rho)$ and arguing as above shows that the tail of (a_k) is bounded below by the terms of a divergent geometric series. □

Theorem 4.60. *Let (a_k) be a sequence of* positive *terms, and assume*

$$\lim_{n\to\infty} \sqrt[n]{|a_n|} = \rho$$

exists. If $\rho < 1$, then (a_k) is absolutely summable. If $\rho > 1$, then (a_k) is not summable.

The proof of the root test is similar to the proof of the ratio test, see Exercise 4.22. In both theorems, nothing can be deduced if the limit fails to exist, or if the limit is 1 (in which case the test is "inconclusive").

Example 4.61 Suppose we wish to test the series

$$\sum_{k=0}^{\infty} \frac{k+1}{2^k} = 1 + \frac{2}{2} + \frac{3}{4} + \frac{4}{8} + \frac{5}{16} + \cdots$$

for convergence; $a_k = (k+1)\,2^{-k}$. Comparison with the geometric series $\sum_k 2^{-k}$ is no help, because the series in question is *larger* than this geometric series. Instead we try the ratio test:

$$\frac{a_{n+1}}{a_n} = \frac{(n+2)\,2^{-(n+1)}}{(n+1)\,2^{-n}} = \frac{(n+2)}{2(n+1)}.$$

This ratio converges to $\rho = 1/2$ as $n \to \infty$, so the series converges absolutely by the ratio test. (Finding the sum of the series is another matter.) □

Example 4.62 Let $p > 0$, and consider the *p-series*

$$\sum_{k=1}^{\infty} \frac{1}{k^p} = 1 + \frac{1}{2^p} + \frac{1}{3^p} + \frac{1}{4^p} + \cdots .$$

None of the tests we have developed so far can resolve the question of convergence for this series; the terms approach 0, so the vanishing criterion is inconclusive. The terms decrease in size more slowly than the terms of an arbitrary convergent geometric series, so there is no obvious comparison to make. Finally, the ratio and root tests both return $\rho = 1$ for every p-series, so these tests are inconclusive as well. □

The next theorem, the *Cauchy test*, is useful for determining convergence of decreasing sequences of positive terms, and will allow us to determine convergence of the p-series.

Theorem 4.63. *Let $(a_k)_{k=1}^{\infty}$ be a sequence of positive terms, and assume $a_{k+1} \leq a_k$ for all $k \geq 1$. If (b_n) is the sequence defined by $b_n = 2^n\, a_{2^n}$, then (a_k) and (b_n) are simultaneously summable or not summable.*

Proof. For $n \in \mathbf{N}$, define the "nth group" of terms in a sequence to be those whose index k is between $2^n + 1$ and 2^{n+1}: The 0th group of (a_k) is $\{a_2\}$, the first group is $\{a_3, a_4\}$, the second group is $\{a_5, \ldots, a_8\}$, the third group is $\{a_9, \ldots, a_{16}\}$, and so forth. There are 2^n terms in the nth group, and all terms but the first two are in exactly one group.

Let c_n denote the sum of the terms in the nth group of (a_k), namely

$$c_n := \sum_{k=2^n+1}^{2^{n+1}} a_k.$$

The sequences $(a_k)_{k=2}^{\infty}$ and $(c_n)_{n=0}^{\infty}$ are simultaneously summable or not (because they each consist of the same terms in the same order).

Since (a_k) is non-increasing,

$$\frac{1}{2} b_{n+1} = 2^n\, a_{2^{n+1}} \leq c_n \leq 2^n\, a_{2^n+1} \leq 2^n\, a_{2^n} = b_n.$$

If (b_n) is summable, then (c_n) is summable by the comparison test, while if (c_n) is summable, then $(b_n/2)$—and thus (b_n) itself—is summable. □

Example 4.64 (The p-series revisited) Fix $p > 0$ and set $a_k = k^{-p}$ for $k \geq 1$. By the Cauchy test, the p-series is convergent iff

$$\sum_{n=1}^{\infty} 2^n \, a_{2^n} = \sum_{n=1}^{\infty} 2^n \, (2^n)^{-p} = \sum_{n=1}^{\infty} (2^{1-p})^n$$

is convergent. This is a geometric series with $r = 2^{1-p}$, and therefore converges iff $p > 1$. Note carefully that this argument proves that the p-series *converges* for $p > 1$; it does *not* say what the sum of the series is, though it does give upper and lower bounds. As of this writing, the sum of the 3-series (the p-series with $p = 3$) is not known exactly, though the value is known to be irrational by a 1978 result of Apéry. By contrast, the value of the $2k$-series (k a positive integer) is known to be a certain rational multiple of π^{2k}. The final result of this book, which relies upon most of the material to come, is the evaluation of the 2-series.

If you are fastidious, you may legitimately complain that we have not defined n^p for non-integer p. This defect is remedied in Chapter 12, after which you may verify that the estimates given here carry over to arbitrary *real* p. □

Alternating Series

Let (a_k) be a sequence of *positive* terms. The series

$$\sum_{k=0}^{\infty} (-1)^k a_k = a_0 - a_1 + a_2 - a_3 + a_4 - a_5 + \cdots$$

is called an *alternating series*.

Alternating series arise frequently when studying power series, and are investigated with different techniques than we have used so far. The idea is to assume that successive terms "tend to cancel" rather than assuming absolute summability. The basic sufficient condition for summability is due to Leibniz and is often called the *alternating series test*:

Theorem 4.65. *Let (a_k) be a sequence of positive terms that decreases to 0, and let A_n be the nth partial sum of the resulting alternating series. Then $A := \lim_n A_n$ exists—the series converges—and*

$$A_{2n-1} < A < A_{2n}$$

for all $n \geq 1$.

Proof. Visually the proof is simplicity itself: The 0th partial sum is a_0, and subsequent partial sums are obtained by moving alternately left and right by smaller and smaller amounts, with the step size going to zero. What can the partial sums do but converge?

A formal proof is based on writing the above argument analytically, considering the even and odd partial sums separately.

Consider the even partial sum A_{2n}. The next even partial sum is

$$A_{2n+2} = A_{2n} - a_{2n+1} + a_{2n+2} = A_{2n} - (a_{2n+1} - a_{2n+2}) < A_{2n}.$$

The proof that the odd partial sums form an increasing sequence is entirely similar. Since each odd partial sum is less than the next even sum,

$$A_{2n-1} < A_{2n+1} < A_{2n+2} < A_{2n} \qquad \text{for all } n > 0. \tag{4.14}$$

Induction on n proves that *every even sum is greater than every odd sum.* In particular, the set of even sums is bounded below, and the set of odd sums is bounded above. Let L be the supremum of the odd partial sums and let U be the infimum of the even partial sums. Since $|U - L| < a_k$ for all k, the squeeze theorem says $U = L$. □

As an immediate consequence, we get a simple, explicit "error bound" that measures the accuracy of estimating the sum of an alternating series by a partial sum: The error is no larger than the size of the first term omitted.

Corollary 4.66. *In the notation of the theorem,* $|A - A_n| < a_{n+1}$.

Example 4.67 If $p > 0$, the sequence $(n^{-p})_{n=1}^{\infty}$ is positive and decreasing, so the Leibniz test implies that the series

$$\sum_{n=1} \frac{(-1)^{n-1}}{n^p} = 1 - \frac{1}{2^p} + \frac{1}{3^p} - \frac{1}{4^p} + \frac{1}{5^p} - \frac{1}{6^p} + \cdots \tag{4.15}$$

converges. As we saw above, this series is absolutely summable iff $p > 1$.

The series with $p = 1$ is the conditionally convergent *alternating harmonic series*,

$$\sum_{n=1} \frac{(-1)^{n-1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \cdots$$

This is a series whose sum S can be found explicitly, see Chapter 14. For the moment, we see from the first three partial sums that

$$0.5 = 1 - \tfrac{1}{2} < S < 1 - \tfrac{1}{2} + \tfrac{1}{3} = 0.83\overline{3}.$$

The alternating harmonic series also has a fairly spectacular rearrangement. Instead of taking positive and negative terms alternately, take one positive and *two* negative terms at a time:

$$\begin{aligned} 1 - \tfrac{1}{2} - \tfrac{1}{4} + \tfrac{1}{3} - \tfrac{1}{6} - \tfrac{1}{8} + \tfrac{1}{5} - \tfrac{1}{10} - \tfrac{1}{12} + \cdots \\ &= \left(1 - \tfrac{1}{2}\right) - \tfrac{1}{4} + \left(\tfrac{1}{3} - \tfrac{1}{6}\right) - \tfrac{1}{8} + \left(\tfrac{1}{5} - \tfrac{1}{10}\right) - \tfrac{1}{12} + \cdots \\ &= \tfrac{1}{2} - \tfrac{1}{4} + \tfrac{1}{6} - \tfrac{1}{8} + \tfrac{1}{10} - \tfrac{1}{12} + \cdots \\ &= \tfrac{1}{2}\left(1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{10} + \tfrac{1}{5} - \tfrac{1}{6} + \cdots\right) = \tfrac{1}{2}S. \end{aligned}$$

Rearrangement of a conditionally convergent series can change the sum! □

The remarkable thing about the rearrangement of the alternating harmonic series is the explicitness of the calculation. The behavior itself is not surprising, in light of the next result. The proof has similarities with the proof of the Leibniz test.

Theorem 4.68. *Let $\sum_k a_k$ be a conditionally summable series. For every real number A, there exists a rearrangement that converges to A. There also exist rearrangements whose partial sums diverge to $+\infty$ or to $-\infty$.*

A formal proof is tedious, but the idea is fairly simple. The sequences of positive and negative terms, (a_k^+) and (a_k^-), both fail to be summable. Suppose $A > 0$. Add up terms from (a_k^+) until the partial sum becomes larger than A; this is guaranteed to happen because the sequences of positive terms is not summable. Now start adding in terms from (a_k^-) until the partial sum becomes smaller than A. Again we use the fact that (a_k^-) is not summable. Repeat *ad infinitum*, selecting positive or negative terms in sequence so that the partial sums bracket the target A.

It is fairly clear that this recipe describes a rearrangement of the original series: Each term appears exactly once in the new sum. Further, the summands approach zero, so the bracketing sums are closer together the longer we carry on. The bracketing sums clearly approach A, so we have the desired rearrangement.

If $A < 0$, start with negative terms, but otherwise use the same idea. To get a rearrangement that diverges to $+\infty$, add up positive terms until the partial sum is larger than 1. Then subtract a single negative term, and add positive terms until the partial sum is larger than 2. Continue in this fashion. You can gauge the accuracy of your intuition about countable sets by the ease with which you see that this procedure works. An untrained person is likely to object that "the positive terms get used up faster than the negative terms"; this is no object, because we only add up finitely many positive terms for each negative term, and there are infinitely many of each!

Another potential snag is that some of the negative terms may be very large in absolute value. However, the sequence of negative terms converges to zero by the vanishing criterion (Proposition 4.51), so after finitely many terms, each negative term is no larger than $1/2$ (say), and each cycle (add a bunch of positive terms and subtract one negative) adds at least $1 - \frac{1}{2}$ to the partial sums, which therefore become arbitrarily large.

Exercises

In all questions where you are asked to find a limit, you should give a complete proof, either with an ε-δ argument, or by citing an appropriate theorem from the text. In questions that have a yes/no answer, give a proof or counterexample, as appropriate.

Exercise 4.1 Are the following true or false?

(a) $\lim_{x\to 0} \frac{1}{x} = +\infty$ (b) $\lim_{x\to 0} \frac{1}{x^2} = +\infty$

(c) $\lim_{x\to 0} \frac{1}{\sqrt{x}} = +\infty$ (d) $\lim_{x\to 0} \frac{1}{\sqrt{|x|}} = +\infty$

In each part, you must prove your answer is correct. ◇

Exercise 4.2 Let $f : (a, b) \to \mathbf{R}$ be a function on an open interval.

(a) Suppose $f(x + h) = f(x) + O(h^2)$ near 0 for all x. Prove that

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = 0 \qquad \text{for all } x \in (a, b).$$

Can you say more?

(b) Suppose there exists a function f' on (a, b) with the property that

$$f(x+h) = f(x) + hf'(x) + o(h) \qquad \text{for all } x \in (a, b).$$

Prove that $f'(x) = \lim_{h\to 0} \dfrac{f(x+h) - f(x)}{h}$ for all x.

(c) Suppose f and g satisfy the condition in part (b). Prove that $f+g$ and fg also satisfy the condition. As a fringe benefit of this calculation, you should find $(f+g)'$ and $(fg)'$ in terms of f, g, f', and g'.

◇

Read the next two questions carefully!

Exercise 4.3 Let $f : A \to \mathbf{R}$ be a function whose domain contains a deleted interval about a. Consider the following condition: For every $\delta > 0$, there exists an $\varepsilon > 0$ such that

$$0 < |x - a| < \varepsilon \implies |f(x) - \ell| < \delta.$$

Is this condition equivalent to "$\lim(f, a) = \ell$"? Give a proof or counterexample. ◇

Exercise 4.4 Let $f : A \to \mathbf{R}$ be a function whose domain contains a deleted interval about a. Consider the following condition: For every $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$0 < |x - a| < \varepsilon \implies |f(x) - \ell| < \delta.$$

Is this condition equivalent to "$\lim(f, a) = \ell$"? Give a proof or counterexample. ◇

Exercise 4.5 This exercise is related to Corollary 4.21.

(a) Prove that if $\lim(f, a) = \ell \neq 0$ and $\lim(g, a)$ does not exist, then $\lim(fg, a)$ does not exist.

(b) Find a pair of functions f and g such that $\lim(f, 0) = 0$ and $\lim(g, 0)$ does not exist, but $\lim(fg, 0)$ exists.

(c) Find a pair of functions f and g such that $\lim(f, 0)$ and $\lim(g, 0)$ do not exist, but $\lim(fg, 0)$ exists.

◇

Exercise 4.6 Define $f : (-1, 1) \to \mathbf{R}$ by $f(x) = 0$ if $x \neq 0$, $f(0) = 1$. Find (with proof) $\lim(f, 0)$, or prove the limit does not exist. ◇

Exercise 4.7 In analogy to the definition of "limit from above" in the text, give a careful definition of "limit from below of f at a." Pay careful attention to the domain of f, as well as the quantified sentence that defines the limit condition. ◇

Exercise 4.8 Suppose $p : \mathbf{R} \to \mathbf{R}$ is a *non-constant* polynomial function. Prove that $\lim(|p|, +\infty) = +\infty$. ◇

Exercise 4.9 Precisely define "$\lim(f, +\infty) = +\infty$." ◇

Exercise 4.10 Precisely define "$\lim(f, x_0) = -\infty$." ◇

Exercise 4.11 Let $f : (0, +\infty) \to \mathbf{R}$ be a function. Prove that

$$\lim_{x \to 0^+} f(\tfrac{1}{x}) = \lim_{x \to +\infty} f(x)$$

in the sense that either both limits exist and are equal, or else neither limit exists. ◇

Exercise 4.12 A set $A \subset \mathbf{R}$ is *dense* if every interval of $\mathbf{R}$ contains a point of A. For example, $\mathbf{Q} \subset \mathbf{R}$ is dense. Suppose f_1 and f_2 are continuous functions on $\mathbf{R}$, and that $f_1|_A = f_2|_A$ for some dense set A. Prove that $f_1 = f_2$ as functions on $\mathbf{R}$.

In other words, a continuous function on A has at most one continuous extension to $\mathbf{R}$. ◇

Exercise 4.13 Let $f : \mathbf{R} \to \mathbf{R}$ be a continuous, non-constant, periodic function. Prove that there exists a smallest positive period. ◇

Exercise 4.14 Let $\mathrm{cb} : \mathbf{R} \to \mathbf{R}$ be the Charlie Brown function.

(a) Does $\lim_{x \to +\infty} x \,\mathrm{cb}(x)$ exist as an extended real number?

(b) Sketch the graph of the function $f : (0, 1) \to \mathbf{R}$ defined by $f(x) = \mathrm{cb}(1/x)$.

(c) Does $\lim(f, 0)$ exist?

◇

Exercise 4.15 If f and g are real-valued functions on $\mathbf{R}$, then $\max(f, g)$ is the function on $\mathbf{R}$ defined by

$$\max(f, g)(x) = \max\big(f(x), g(x)\big) = \begin{cases} g(x) \text{ if } f(x) \leq g(x) \\ f(x) \text{ if } g(x) \leq f(x) \end{cases}$$

and $\min(f, g)$ is defined similarly.

(a) Prove that if f and g are continuous on $\mathbf{R}$, then $\max(f, g)$ and $\min(f, g)$ are continuous on $\mathbf{R}$. Suggestion: Use Theorem 2.24.

(b) Prove that every continuous function $f : \mathbf{R} \to \mathbf{R}$ may be written as a difference of continuous, *non-negative* functions, say $f = f_+ - f_-$. Part (a) and a sketch may help.

◇

Exercise 4.16 Each part should be interpreted in the sense of equation (4.6).

(a) Prove that $(+\infty) + x = +\infty$ for all $x \in \mathbf{R}$.

(b) Prove that $(+\infty) + (+\infty) = +\infty$.

(c) Prove that if $\ell > 0$, then $-\ell \cdot (+\infty) = -\infty$.

◇

Exercise 4.17 Each part should be interpreted in the sense of projective infinite limits.

(a) Prove that $\infty + x = \infty$ for all $x \in \mathbf{R}$.

(b) Prove that $\infty + \infty$ is indeterminate.

(c) Prove that $x/0 = \infty$ for all $x \neq 0$.

◇

Exercise 4.18 Let $p : \mathbf{RP}^1 \to S^1$ be stereographic projection, see Exercise 3.7.

(a) Show that $\lim_{t \to \infty} p(t) = (0, 1)$ in the sense that each component function has the advertised limit.

(b) Prove that the rational function $t \mapsto 1/t$ corresponds under p to reflection of the unit circle in the horizontal axis.

(c) Prove that the rational function $t \mapsto \frac{t-1}{t+1}$ corresponds under p to the mapping $(x, y) \mapsto (-y, x)$, namely to rotation of the unit circle by a quarter-turn counterclockwise about the origin.

◇

Exercise 4.19 Prove that under the hypotheses of Theorem 4.30,

$$\ell^+ := \lim(f, x^+) = \inf\{f(y) \mid x < y < b\}.$$

Given an example of a non-decreasing function $f : [0, 1] \to \mathbf{R}$ that has infinitely many discontinuities. ◇

Exercise 4.20 Prove Corollary 4.40. ◇

Exercise 4.21 Prove Theorem 4.41. ◇

Exercise 4.22 Prove Theorem 4.60 (the root test). ◇

Exercise 4.23 This problem refers to Example 4.47. Write $a = \sqrt{b}$.

(a) Prove that $x_n^2 > b$ for all $n \in \mathbf{N}$, and that the sequence (x_n) is decreasing. Conclude that (x_n) is bounded below by a.

(b) For $n \in \mathbf{N}$, let $e_n = x_n - a$ be the error in estimating $\sqrt{b}$ by x_n. Prove that

$$e_{n+1} = \frac{e_n^2}{2x_n} < \frac{e_n^2}{2a} \qquad \text{for all } n \in \mathbf{N},$$

and hence that

$$e_{n+1} < 2a\left(\frac{e_1}{2a}\right)^{2^n} \qquad \text{for all } n \in \mathbf{N}. \tag{4.16}$$

This says the number of decimals of accuracy grows exponentially with each iteration.

(c) Let $b = 3$ and $x_0 = 2$. Prove (without evaluating $\sqrt{3}$ numerically of course) that $e_1/(2a) < 1/10$, and conclude that the sixth term approximates $\sqrt{3}$ to 31 decimal places, i.e., that $|x_6 - \sqrt{3}| < 5 \cdot 10^{-32}$. Suggestion: By arithmetic, $1.7 < \sqrt{3} < 1.8$.

◇

Exercise 4.24 By Example 4.47, there is a function $\sqrt{\ } : [0,\infty) \to [0,\infty)$ having the property that $(\sqrt{x})^2 = x$ for all $x \geq 0$. Prove that $\sqrt{\ }$ is continuous. Suggestion: First show $\sqrt{\ }$ is increasing, then use the identity $(\sqrt{x}-\sqrt{a})(\sqrt{x}+\sqrt{a}) = x-a$. ◇

Exercise 4.25 Find each of the following limits or prove it does not exist.

(a) $\lim\limits_{x\to -1} \dfrac{x^2-1}{x+1}$

(b) $\lim\limits_{x\to 1} \dfrac{1-\sqrt{x}}{1-x}$

(c) $\lim\limits_{x\to 0} \dfrac{1-\sqrt{1-x^2}}{x}$

(d) $\lim\limits_{x\to 0} \dfrac{1-\sqrt{1-x^2}}{x^2}$

(e) $\lim\limits_{x\to \infty} \sqrt{x^2+x}-x$

You may use the result of Exercise 4.24. ◇

Exercise 4.26 Define $f : (0,\infty) \to (0,\infty)$ by $f(x) = \sqrt{2+x}$.

(a) Prove that f is continuous, and find the fixed point(s) of f.

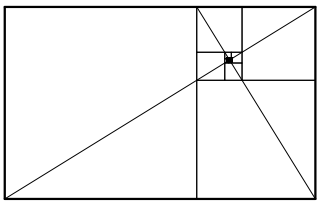
(b) Define $(x_n)_{n=0}^{\infty}$ by $x_0 = \sqrt{2}$, $x_{n+1} = f(x_n)$ for $n \geq 1$; that is,

$$x_1 = \sqrt{2+\sqrt{2}}, \qquad x_2 = \sqrt{2+\sqrt{2+\sqrt{2}}}, \qquad \ldots$$

Prove that (x_n) converges, and find the limit.

The limit is denoted $\sqrt{2+\sqrt{2+\sqrt{2+\cdots}}}$. ◇

Exercise 4.27 The Ancient Greeks believed that the most aesthetically pleasing rectangle is a *golden rectangle*, one that keeps the same proportions when a square is removed:



(a) Find the ratio τ of width to height for a golden rectangle. Compute τ^2, $\tau - 1$, and $1/\tau$.

(b) Show that

$$\tau = [1; 1, 1, 1, \ldots] = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \ddots}}$$

(There are a few ways to proceed.)

(c) Interpret parts (a) and (b) in terms of the right-hand diagram.

◇

Exercise 4.28 Evaluate the infinite continued fraction $[2; 2, 2, 2, \ldots]$.
◇

Exercise 4.29 Let (a_k) be a sequence of real or complex numbers, and let $b_k = a_{k+1} - a_k$ be the sequence of "differences".

(a) Use induction on n to prove that

$$\sum_{k=0}^{n} b_k = a_{n+1} - a_0 \qquad \text{for all } n \in \mathbf{N}.$$

A sum of this form is said to be *telescoping*.

(b) Suppose $\lim_k a_k = \ell$ exists; prove that (b_k) is summable, and find the sum of the series.

(c) Evaluate the infinite sum $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k^2 + k} = \sum_{k=1}^{\infty} \Big(\frac{1}{k} - \frac{1}{k+1} \Big)$.

◇

Exercise 4.30 Evaluate the series

$$\sum_{n=1}^{\infty} \frac{1}{(2n)^2} = \frac{1}{4} + \frac{1}{16} + \frac{1}{36} + \frac{1}{64} + \cdots$$

$$\sum_{n=1}^{\infty} \frac{1}{(2n+1)^2} = \frac{1}{1} + \frac{1}{9} + \frac{1}{25} + \frac{1}{49} + \cdots$$

in terms of $S = \displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{1}{1} + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \frac{1}{36} + \frac{1}{49} + \cdots$.
◇

Exercise 4.31 Generalize the preceding exercise: Let $p > 1$, so that the p-series $\sum_n n^{-p}$ is absolutely summable, and let S be the sum of the series. Evaluate the sum of the even terms and the sum of the odd terms. ◇

Exercise 4.32 Suppose $f : (a, b) \to \mathbf{R}$ is continuous, and that $(x_n)_{n=0}^{\infty}$ is a Cauchy sequence in (a, b). Is the sequence $\big(f(x_n)\big)$ necessarily Cauchy? Reassurance: Deciding whether the answer is "yes" or "no" is likely to be the most difficult part of the problem. ◇

Chapter 5

Continuity on Intervals

Throughout the chapter, $f : [a, b] \to \mathbf{R}$ is a continuous function, that is, f is continuous at x for each $x \in [a, b]$. Continuity of a function at a point is a local property, and even continuity of a function on an interval is, on the surface, nothing more than a collection of local conditions. In this chapter we deduce some truly global properties (such as boundedness) of continuous functions on a closed, bounded interval $[a, b]$. These results are the technical foundation stones of calculus, and while some of their statements are "intuitively obvious," all of them require the completeness axiom of $\mathbf{R}$ in a fundamental way.

A theme runs throughout this chapter, resting on the assumption that the domain of f is a closed, bounded interval. Suppose we wish to show f satisfies some property $\mathcal{P}$ (such as "boundedness") on $[a, b]$. By the nature of $\mathcal{P}$ we know that a continuous function has the property locally (in an interval about each point). We start at a, where f has the property. By continuity, the property holds on some interval $[a, t)$ with $t > a$. Now consider the supremum of t such that f satisfies $\mathcal{P}$ on $[a, t)$. If $t < b$ we arrive at a contradiction, since continuity at t implies $\mathcal{P}$ holds on a slightly larger interval $[a, t')$; thus $t = b$, and $\mathcal{P}$ holds on $[a, b)$. By continuity of f at b, the property holds on the entire interval $[a, b]$. In this very rough sketch, each of the three conditions "closed, bounded interval" is used in an essential way. If even one of these hypotheses is omitted, or if the function f is discontinuous at even one point, then f no longer has the property $\mathcal{P}$ in general.

It was not until the early 20th Century that mathematicians found axiomatic criteria to replace the condition of being a "closed, bounded interval." Such criteria are needed to study functions of several variables, where "intervals" make no sense. The abstract conditions are

called "compactness" (replacing closed and bounded) and "connectedness" (replacing interval). Each of these conditions can be expressed in terms of a game, as we did for continuity in Chapter 4, but the work required to explain the rules of the game, and to see why the general criteria are the "correct" ones, would take us quite far afield. A thorough study of these "topological" conditions belongs in a more advanced course, and they will not be mentioned again.

5.1 Uniform Continuity

We saw that the equation $\ell = \lim(f, a)$ can be viewed as a two-person game. In this section, we meet a "stronger" version of continuity that cannot be phrased naturally in o notation. However, there *is* a natural game-theoretic interpretation. Remember that $f : [a, b] \to \mathbf{R}$ is assumed to be continuous at every point of its domain.

The hypothetical adversaries, Players ε and δ, are looking for variety in their game, because ε keeps losing. Recall that they have been given a function f, and they have agreed to take $\ell = f(x)$ when playing the continuity game at $x \in [a, b]$. The function f they are using gives Player δ a winning strategy at every $x \in [a, b]$. This means that they fix an x, Player ε chooses a tolerance $\varepsilon > 0$, and Player δ successfully "meets" the tolerance. However, according to the rules, the point x is specified *in advance*, before either ε or δ is chosen. Player δ chooses with knowledge of both ε and x. In an attempt to make the game more difficult for Player δ, they change the rules as follows:

The function f is given. Player ε chooses a tolerance. Now Player δ is required to meet the tolerance with a *single* δ that works *for all* x. If Player δ has a winning strategy (as before, against a perfect player), then the function f is "uniformly continuous." There is no need to assume the domain of f is an interval, much less a closed, bounded interval:

Definition 5.1 Let $f : X \to \mathbf{R}$ be a function. We say f is *uniformly continuous* (on X) if, for every $\varepsilon > 0$, there is a $\delta > 0$ such that if $x, y \in X$ and $|x - y| < \delta$, then $|f(x) - f(y)| < \varepsilon$.

Uniform continuity is a "global" property: It requires simultaneous consideration of all points of X. (This is why o notation is not well-suited.) Even if f is "well-behaved" (say constant) in some neighborhood of each point of X, it does not follow that f is uniformly continuous on X, see Example 5.3.

If f is uniformly continuous on X, then *a fortiori* the restriction of f to a non-empty subset is also uniformly continuous. Clearly, a uniformly continuous function is continuous at each point of its domain, since Player δ can be lazy and use the same δ regardless of x. The best way to understand the difference between continuity on X and uniform continuity on X is to consider examples. Lemma 5.2 gives a necessary (but not sufficient) criterion for uniform continuity.

Lemma 5.2. *Let $f : X \to \mathbf{R}$ be a uniformly continuous function on a bounded interval. Then f is a bounded function: There exists an $M > 0$ such that $f = A(M)$ on X.*

Proof. The idea is that by uniform continuity, there exists a $\delta > 0$ such that f varies by no more than 1 on every interval of length δ. Because X is bounded, it can be covered by finitely many intervals of length δ, so the total variation of f is finite.

Formally, take $\varepsilon = 1$; by uniform continuity, there is a $\delta > 0$ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < 1$ (provided x and y are in X). Let x_0 be the midpoint of X, and let ℓ denote the radius of X. Pick a natural number N larger than ℓ/δ. The idea is that every point $x \in X$ can be "joined" to x_0 by a "chain" of N overlapping intervals of length less than δ; on each such interval, the function values vary by at most 1, so by the triangle inequality the function value $f(x)$ differs from $f(x_0)$ by at most N, regardless of x.

For the record, here is a complete argument. Let $x \in X$, and consider the finite sequence $(x_i)_{i=0}^N$ of equally spaced points that joins x_0 to $x = x_N$. There is even a simple formula for the ith point:

$$x_i = x_0 + \tfrac{i}{N}(x - x_0) = \left(1 - \tfrac{i}{N}\right) x_0 + \tfrac{i}{N}\, x, \qquad 0 \le i \le N.$$

For each i, $|x_i - x_{i-1}| = |x - x_0|/N < \ell/N < \delta$, so $|f(x_i) - f(x_{i-1})| < 1$. By the triangle inequality,

$$|f(x) - f(x_0)| = \left| \sum_{i=1}^{N} \Big(f(x_i) - f(x_{i-1}) \Big) \right| \le \sum_{i=1}^{N} \Big| f(x_i) - f(x_{i-1}) \Big| < N.$$

This estimate holds for all $x \in X$, so again by the triangle inequality,

$$|f(x)| \le |f(x) - f(x_0)| + |f(x_0)| < N + |f(x_0)| =: M$$

for all $x \in X$. □

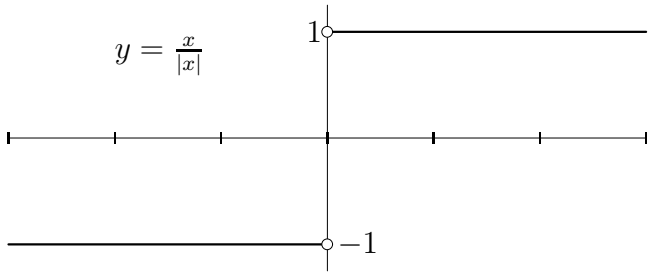


Figure 5.1: A locally constant function that is not uniformly continuous.

Example 5.3 The identity function $I : \mathbf{R} \to \mathbf{R}$ is uniformly continuous; taking $\delta = \varepsilon$ (independently of x) gives a winning strategy. Though I is not bounded on $\mathbf{R}$, there is no conflict with Lemma 5.2 since $\mathbf{R}$ is not a bounded interval. In accord with the lemma, the restriction of I to a bounded interval *is* bounded.

The reciprocal function $f : (0,1) \to \mathbf{R}$ is continuous on $(0,1)$ but is not bounded, hence is not uniformly continuous by Lemma 5.2, see also Figure 5.2.

The function $g(x) = x/|x|$ (the restriction of the signum function to $\mathbf{R}^\times$, Figure 5.1) is locally constant (for every $a \neq 0$, there is an open interval about a on which g is constant), but is *not* uniformly continuous! Take $\varepsilon = 1$; no matter how small $\delta > 0$ is, the points $x = -\delta/3$ and $y = \delta/3$ are δ-close, but $|g(x) - g(y)| = 2 > 1 = \varepsilon$. This example emphasizes the global nature of uniform continuity, and shows that a bounded function can fail to be uniformly continuous. If you are tempted to protest that g is not continuous at 0, remember that "continuity" makes no sense at points not in the domain of g. □

Uniform continuity has a geometric interpretation: The image of an interval can be made arbitrarily short by taking the interval sufficiently short. Precisely, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that the image of an arbitrary interval of length δ is contained in some interval of length ε. Reconsider the examples above in light of this observation. An interval of length δ has the form $(a, a+\delta)$; under the reciprocal function, such an interval can have arbitrarily long image, see Figure 5.2. Under the signum function, if the interval straddles the origin the image is the two-point set $\{-1, 1\}$, so the image is not contained in an interval of length less than 2. By contrast, the image of $(a, a+\delta)$ under the

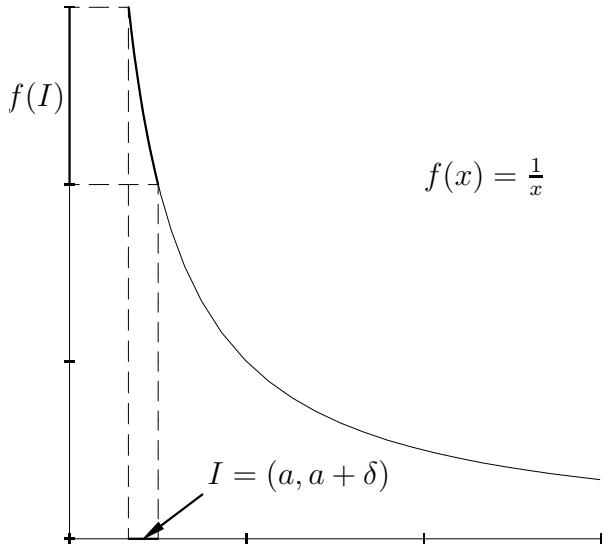


Figure 5.2: The reciprocal maps arbitrarily short intervals to "long" intervals.

identity function is the interval $(a, a+\delta)$, which tautologically can be made arbitrarily small by taking δ arbitrarily small!

The main result of this section, Theorem 5.5 below, is a simple criterion for uniform continuity that is of both theoretical and practical importance. To reduce "uniform continuity" to a more manageable form, we introduce an auxiliary criterion. For a fixed $\varepsilon > 0$, we say the function f is "ε-tame on X" if there exists a $\delta > 0$ such that

$$x,\ y \in X \text{ and } |x - y| < \delta \quad \Longrightarrow \quad |f(x) - f(y)| \leq \varepsilon.$$

For example, if $f = A(M)$ on X, then f is $2M$-tame on X by the triangle inequality. Similarly, every characteristic function is 1-tame. Uniform continuity on X is equivalent to "being ε-tame on X for every $\varepsilon > 0$."

The proof of Lemma 5.2 shows that an ε-tame function on a bounded interval is bounded. Consequently, though the reciprocal function is continuous, it is *not* ε-tame, no matter how large ε is.

The condition of being ε-tame satisfies a "patching" property:

Lemma 5.4. *Suppose f is ε-tame on the closed intervals $[a, t]$ and $[t, c]$, and that f is continuous at t. Then f is ε-tame on the union $[a, c]$.*

Proof. Use continuity of f at t to choose $\delta_1 > 0$ so that $|x - t| < \delta_1$ implies $|f(x) - f(t)| < \varepsilon/2$. The triangle inequality implies f is ε-tame on $(t - \delta_1, t + \delta_1)$. Next choose $\delta_2 > 0$ and $\delta_3 > 0$ that "work" on $[a, t]$ and $[t, c]$, respectively, and set $\delta = \min(\delta_1, \delta_2, \delta_3) > 0$.

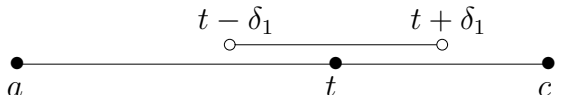


Figure 5.3: Patching intervals on which f is ε-tame.

A glance at Figure 5.3 shows that if x and y are points in $[a, c]$ with $|x - y| < \delta$, then *both points* lie in *one* of the three intervals $[a, t]$, $[t, c]$, or $(t - \delta_1, t + \delta_1)$. Consequently f is ε-tame on $[a, c]$. □

Theorem 5.5. *If $f : [a, b] \to \mathbf{R}$ is a continuous function on a closed, bounded interval, then f is uniformly continuous.*

Proof. We will show that under the hypotheses, f is ε-tame on $[a, b]$ for every $\varepsilon > 0$. Fix $\varepsilon > 0$, and consider the set

$$B = \{t \in (a, b] \mid f \text{ is } \varepsilon\text{-tame on } [a, t]\}.$$

Our goal is to show that $b \in B$. This is accomplished by the sketch in the introduction of this chapter, interplaying continuity of f, the completeness axiom of $\mathbf{R}$, and the fact that $[a, b]$ is a closed, bounded interval.

Because f is continuous at a, there exists a $\delta > 0$ such that $f = A(\varepsilon/2)$ on $[a, a+2\delta)$. By the triangle inequality, f is ε-tame on $[a, a+\delta]$. Thus $a+\delta \in B$, so the set B is non-empty; clearly, B is bounded above (by b), so by completeness B has a supremum, say β.

We claim that $\beta = b$; to see this, suppose f is ε-tame on $[a, \alpha)$ with $\alpha < b$. By continuity of f at α, there is a $\delta > 0$ such that f is ε-tame on $[\alpha - \delta, \alpha + \delta]$, while f is ε-tame on $[a, \alpha - \delta]$ by assumption. Lemma 5.4 implies f is ε-tame on $[a, \alpha + \delta]$, proving that α is not an upper bound of B. This is the contrapositive of "$\sup B = b$."

We still do not know $b \in B$; it could be that $B = [a, b)$. However, f is continuous at b, hence ε-tame on some interval $[b - \delta, b]$, and the previous paragraph shows that f is ε-tame on *every* interval $[a, \alpha]$ with $\alpha < b$, in particular on $[a, b - \delta]$. Another application of Lemma 5.4 proves f is ε-tame on $[a, b]$. Since $\varepsilon > 0$ was arbitrary, the theorem is proved. □

Theorem 5.5 implies, for instance, that the restriction of a rational function p/q to an interval $[a, b]$ is uniformly continuous (provided q is non-vanishing in $[a, b]$). Thus the reciprocal function $x \mapsto 1/x$ is uniformly continuous on $[\delta, 1]$ for each $\delta > 0$. As noted above, the restriction to $(0, 1]$ is *not* uniformly continuous. Uniform continuity of f is as much a property of the domain as it is a property of the "rule" that defines f.

5.2 Extrema of Continuous Functions

Theorem 5.5 has a technical consequence that is so important it deserves the title of "theorem" rather than "corollary." Theorem 5.6 is called the *Extreme Value Theorem*.

Theorem 5.6. *Let $f : [a, b] \to \mathbf{R}$ be a continuous function. Then there exist points $x_{\min}$ and $x_{\max} \in [a, b]$ such that*

$$f(x_{\min}) \leq f(x) \leq f(x_{\max}) \qquad \text{for all } x \in [a, b].$$

This theorem does *not* assert that the points $x_{\min}$ and $x_{\max}$ are uniquely determined; in the "extreme" case f is a constant function and *every* point of the interval is *both* $x_{\min}$ and $x_{\max}$. The theorem also gives no information as to *where* in the domain these points may be; other tools must be used for this purpose. Finally, the Extreme Value Theorem says nothing about functions with even a single discontinuity, nor about functions whose domain is not a closed, bounded interval of real numbers. What the theorem asserts is the existence—in a certain infinite set of numbers (the set of values of f)—of a *largest* number and a *smallest* number. This is better than knowing the set of values is bounded above and below, which is already quite useful information; it is a "hunting license" for maxima and minima of functions, a guarantee that under suitable hypotheses, the quarry actually exists.

Proof. By Theorem 5.5, f is uniformly continuous on $[a, b]$, hence f is bounded on $[a, b]$ by Lemma 5.2. Consider the image $f([a, b])$; it is certainly non-empty, and as just observed is bounded, both above and below. By completeness, the image has a supremum $y_{\sup}$ and an infimum $y_{\inf}$. We want to show these numbers are function values. This will be done by proving that if $y_{\sup}$ is *not* a function value, then Theorem 5.5 is false.

If $y_{\sup}$ is not a function value, then $f(x) < y_{\sup}$ for every $x \in [a, b]$ (note the strict inequality). Consider the function $g : [a, b] \to \mathbf{R}$ defined by

$$g(x) = \frac{1}{y_{\sup} - f(x)}.$$

By hypothesis the denominator is non-vanishing on $[a, b]$, so by continuity of f and Theorem 4.19 the function g is continuous. However, the denominator of g can be made arbitrarily small, since by definition of supremum, for every $\varepsilon > 0$ there is an x_0 with $y_{\sup} - \varepsilon < f(x_0)$, that is, $0 < y_{\sup} - f(x_0) < \varepsilon$. This implies g is not bounded above, contradicting Theorem 5.5 as desired. A similar argument using the function $h(x) = 1/(f(x) - y_{\inf})$ shows f achieves a minimum value. □

In calculus courses, the extreme value theorem is usually stated without proof, or at best with a plausibility argument. One must be wary of putting credence in a plausibility argument, for the following reason: The statement of the extreme value theorem makes sense for functions $f : \mathbf{Q} \to \mathbf{Q}$, but while any plausibility argument is likely to apply to such a function, the conclusion of the theorem is false. Examples are given below.

A typical plausibility argument is based on the "definition" that a "continuous function" is one whose graph can be drawn without picking up your pencil.[1] Because the domain of f is an interval that contains its endpoints, the graph of f must "start" somewhere on the left and "end" somewhere on the right. The pencil must "therefore" reach a highest point and a lowest point; the horizontal locations of these points are $x_{\max}$ and $x_{\min}$.

5.3 Continuous Functions and Intermediate Values

The other fundamental property of a continuous function f on a closed, bounded interval is the "intermediate value property." In simplest form, this says that if f is negative somewhere and positive somewhere (else), then f must be zero somewhere. Generally, the image of an interval of real numbers under a continuous function is an interval. To put this property in context, we make a brief historical digression.

[1] For many common functions this is true enough, though it differs substantially from Definition 4.36. How many differences can you name?

Deficiencies of the Rational Numbers in Analysis

In terms of the heuristic definition of continuity, the intermediate value property is "obvious": If the graph of f starts below the x-axis and ends above the x-axis, then the graph must cross the axis somewhere. But under scrutiny, this argument is shaky. "Surely it is inconceivable that a graph could go from below the axis to above without crossing" sounds like wishful thinking. Indeed, the misperception that this fact is obvious goes back at least to Euclid, who postulated that every line through the center of a circle intersects the circle in two antipodal points. Suppose the plane consists of all points with *rational coordinates*, that is, $\mathbf{Q}\times\mathbf{Q}$. (Without the hindsight that "a plane is $\mathbf{R}\times\mathbf{R}$" there is no good reason to assume a "plane" is *not* $\mathbf{Q}\times\mathbf{Q}$. To our eyes, they look the same.) The circle of radius 1 centered at the origin has equation $x^2+y^2=1$. This equation makes sense over the rational numbers, and has infinitely many rational solutions. The line with equation $y=x$ also "lives" in the rational plane, and passes through the center of the circle (see Figure 5.4 for a stylized depiction). Solving these equations simultaneously gives the points of intersection as $\pm(x,x)$, where $x^2=1/2$. But there is no rational number with this property, so this line *does not intersect the circle at all*! However, the rational plane $\mathbf{Q}\times\mathbf{Q}$ satisfies all of Euclid's axioms, so his "self-evident" postulate that certain lines and circles intersect is *not generally true*.

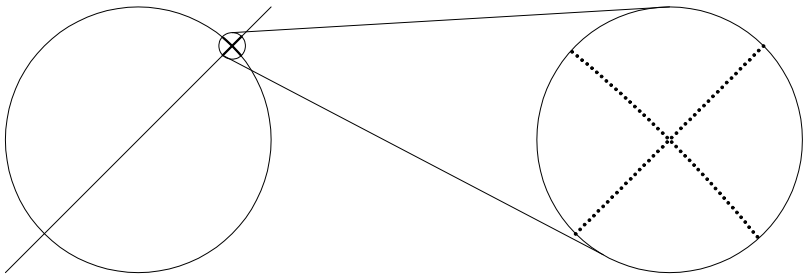


Figure 5.4: Does a circle intersect every line through its center?

This deficiency can be formulated in terms of functions, showing that the heuristic argument for the intermediate value property is incomplete. Suppose we take the number line to be $\mathbf{Q}$, and consider the polynomial function $f:\mathbf{Q}\to\mathbf{Q}$ defined by $f(x)=x^2-2$. The same estimates used in Chapter 4 show that this function is continuous on $\mathbf{Q}$. At $x=0$, the graph lies below the x-axis, and at $x=2$ the graph lies

above. However, the graph never hits the axis on the closed, bounded *rational* interval $[0, 2]$, because $x^2 - 2 \neq 0$ for all rational x. This is not an isolated example; a general polynomial with rational coefficients will be positive somewhere, negative somewhere, but zero nowhere.

These observations are perhaps even more striking for rational functions, such as $g = 1/f : \mathbf{Q} \to \mathbf{Q}$ defined by $g(x) = 1/(x^2 - 2)$. The domain and range are correct; the expression on the right is defined for every rational x, and is itself rational. This function is continuous on $\mathbf{Q}$ because it is a quotient of polynomials, and the denominator is nowhere vanishing! However, g is not uniformly continuous on $[0, 2]$, since it is unbounded on this interval.

You might be tempted to argue, "Yes, but $\mathbf{R}$ has no gaps, so this cannot happen in $\mathbf{R}$." But what *is* a gap, and how do we know $\mathbf{R}$ has none? To emphasize, the theorems of this chapter, as well as our intuition about continuous functions, are predicated on using intervals of *real numbers*, and this intuition is based on the completeness property of $\mathbf{R}$. Plausibility arguments can be incorrect or seriously misleading when applied carelessly; only with precise definitions and logical, deductive proof can one be sure of avoiding errors.

These remarks should cause you to question your intuition. Lest you come to feel everything you know is unjustifiable, let us quickly review the relevant facts. Intervals and absolute value make sense in every ordered field, including both $\mathbf{Q}$ and $\mathbf{R}$. Consequently, we may speak meaningfully of limit points and (uniform) continuity of functions even if our number line is $\mathbf{Q}$. The proof of Lemma 5.2, that uniformly continuous functions are bounded, uses nothing but properties of ordered fields, and therefore also holds in $\mathbf{Q}$. However, in proving Theorem 5.5, we defined a certain set, then took its supremum. Similarly, we used suprema in proving the extreme value theorem. We therefore have no reason to expect the analogues of these theorems to hold if our number line is $\mathbf{Q}$. This pessimism is justified by the counterexamples discussed above: The "extreme value theorem in $\mathbf{Q}$" is simply not true. Further, we have seen that a continuous function on $\mathbf{Q}$ can take both positive and negative values without vanishing anywhere, contrary to our intuition. Our next goal is to prove that our intuition about $\mathbf{R}$ is correct: A continuous function on $\mathbf{R}$ that achieves both positive and negative values must have a zero. This conclusion will have a multitude of useful consequences; for example, it will follow that every positive real number has a cube root, fourth root, and generally a radical of every integer order.

The Intermediate Value Theorem

The next definition makes precise the idea that "if f achieves two values, then it achieves every value in between." Notice how simply the criterion is expressed in terms of intervals.

Definition 5.7 Let f be a function. We say f has the *intermediate value property* if, for every interval I contained in the domain of f, the image $f(I)$ of I is also an interval.

To say the same thing differently, if $[a, b]$ is contained in the domain of f and if $f(a) \neq f(b)$, then for every c between $f(a)$ and $f(b)$, there exists an $x \in (a, b)$ with $f(x) = c$. You should convince yourself this condition is logically equivalent to Definition 5.7. Our geometric intuition suggests that continuous functions (on intervals) *do* have the intermediate value property. In fact they do, by the *Intermediate Value Theorem*:

Theorem 5.8. *Let f be a continuous function whose domain is an interval of real numbers. Then f has the intermediate value property.*

The intermediate value theorem is a "hunting license" in the same sense as Theorem 5.6. Rather than hunting for extrema, we now seek solutions of the equation $f(x) = c$, subject to $a \leq x \leq b$. Theorem 5.8 says that if f is a continuous function on $[a, b]$, and if the number c is between $f(a)$ and $f(b)$, then the equation $f(x) = c$ *has a solution* $x \in (a, b)$. The theorem does not say there is only one solution, and does not give any information about the location of the solution(s). As with the extreme value theorem, these matters must be investigated with other tools.

Proof. Because the restriction of a continuous function is continuous, it suffices to prove that if f is continuous on $[a, b]$, and if $f(a) < c < f(b)$, then there is an $x_0 \in (a, b)$ with $f(x_0) = c$. (The analogous claim with the inequality reversed then follows by consideration of $-f$.) In fact, it suffices to assume $c = 0$, since f can be replaced by $f - c$. The intuitive idea is to "watch the pencil to see when the tip crosses the x-axis". Of course, such reasoning would be circular, since we are trying to *prove the pencil crosses the axis*. Instead, we seek the "largest" t such that f is negative on $[a, t]$. Now, there is no such t, but there *is* a supremum; this is how completeness of $\mathbf{R}$ enters the picture.

Recall the contrapositive of Theorem 4.23: If $\lim(f, a) < 0$, then $f < 0$ on some deleted interval about a (namely, there is a $\delta > 0$ such

that $f(x) < 0$ for $0 < |x - a| < \delta$). The analogous assertion with a positive limit is also true.

Assume $f(a) < 0 < f(b)$, and consider the set

$$B = \{x \in (a, b) \mid f(t) < 0 \text{ for all } t \in [a, x]\}.$$

Since $\lim(f, a^+) = f(a) < 0$, there is a $\delta > 0$ such that $f < 0$ on $[a, a + \delta]$; thus $a + \delta \in B$, so B is non-empty. At the other endpoint, $\lim(f, b^-) = f(b) > 0$, so there is a $\delta' > 0$ such that $f > 0$ on the interval $[b - \delta', b]$. This means "$f(t) < 0$ for all $t \in [a, b - \delta']$" is *false*, or $b - \delta' \notin B$. Thus B is bounded above by $b - \delta'$.

Let $x_0 = \sup B$; then $x_0 \leq b - \delta' < b$, and

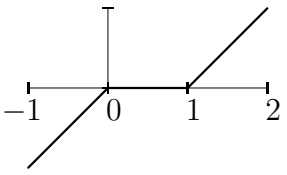
$$(*) \qquad f < 0 \text{ on } [a, x] \text{ for every } x < x_0.$$

The claim is that $f(x_0) = 0$. To prove this, we show that $f(x_0) \neq 0$ implies $x_0 \neq \sup B$.

If $f(x_0) > 0$, then $f > 0$ on some open interval about x_0 by continuity of f, contradicting $(*)$. If, on the other hand, $f(x_0) < 0$, then $f < 0$ on some open interval about x_0, so by $(*)$ $f < 0$ on some interval $[a, x]$ with $x > x_0$, implying x_0 is not an upper bound of B. The remaining possibility, $f(x_0) = 0$, must therefore hold, which completes the proof. □

This proof finds the smallest solution of $f(x) = c$ in the interval $[a, b]$; in particular, *there is a smallest solution.* There is also a largest solution, found by an obvious modification of the argument. Generally there is no "second-smallest" solution. The function $f : [-1, 2] \to \mathbf{R}$ defined by

$$f(x) = \begin{cases} x & \text{if } -1 \leq x < 0 \\ 0 & \text{if } 0 \leq x \leq 1 \\ x - 1 & \text{if } 1 < x \leq 2 \end{cases}$$



has a first zero and a last zero, but no "second" zero.

5.4 Applications

The intermediate value theorem can be used to prove several interesting things about numbers and functions. The sampling given here includes

existence of nth roots of positive reals, existence of a real root for every polynomial of odd degree, and the *bisection method*, a numerical algorithm for approximating solutions of $f(x) = 0$ for a continuous function f.

Radicals of Real Numbers

Let $n \geq 2$ be an integer. By definition, an *nth root* of a number c is a number t (usually in the same field) with $t^n = c$. We now have enough tools at our disposal to prove that every positive real number has a unique positive nth root. The situation is more haphazard for negative numbers because of the prejudice of considering only *real* numbers. A more satisfactory picture emerges when working with complex numbers: Every non-zero complex number has exactly n distinct (complex) nth roots. As we shall see in Chapter 15, these roots lie at the vertices of a regular n-gon centered at 0 in the complex plane.

The intermediate value theorem almost immediately implies existence of nth roots of positive real numbers. The trick is to concoct a function whose zero is the desired radical.

Theorem 5.9. *Let $c \in \mathbf{R}$ be positive, and let $n \geq 2$ be an integer. There exists a unique positive real number t with $t^n = c$.*

Proof. Uniqueness is elementary: If $0 < a < b$, then $0 < a^n < b^n$ for every positive integer n, so $a^n = c$ can hold for at most one $a > 0$. What must be shown is *existence* of an nth root. Define $f : [0, \infty) \to \mathbf{R}$ by $f(x) = x^n - c$; we wish to show that f has a positive root. Since f is a polynomial function, f is continuous on $[0, \infty)$. By direct evaluation, $f(0) = -c < 0$, so it suffices to show there is an $x > 0$ with $f(x) > 0$. If $c > 1$, then take $x = c$:

$$f(c) = c^n - c = c\,(c^{n-1} - 1) > 0,$$

and we are done. If $0 < c < 1$, then $c^n < 1$, so $f(1) = 1 - c^n > 0$, and again we are done. If $c = 1$ the conclusion is obvious. □

It is gratifying to see how far we have come. Aside from the details in the construction of $\mathbf{R}$, everything in this corollary has been built on set theory and logic. You should also appreciate how much abstract machinery is required to describe an elementary geometric concept—the diagonal of a unit square—in terms of numbers (let alone sets). Of

course, in order to be impressed you must admit how much of school mathematics is based upon unproved assertions.

Customarily one writes $t = c^{1/n}$ or $t = \sqrt[n]{c}$ for the nth root of c. With the exponential notation, the rules

$$a^{x+y} = a^x \cdot a^y \quad \text{and} \quad a^{xy} = (a^x)^y$$

hold for all *rational* exponents. Until now, we only had suitable definitions of exponentiation for *integer* exponents, and we have still not defined exponentiation with non-rational, or non-real, exponents.

Real Roots of Polynomials

Finding roots of polynomials is one of the oldest problems of mathematics; over 4000 years ago, the Babylonians knew how to solve what we would call the general quadratic equation. Analogous formulas for cubic and quartic equations were not found until the 15th Century. In the early 19th Century, N. H. Abel showed that there *does not exist* a "quintic formula," in the sense that the roots of a general polynomial of degree five or more cannot generally be written as radical expressions in the coefficients.

Even for cubic and quartic polynomials, the algebraic formulas are messy, and it is desirable to have simpler (if less precise) tools for gleaning information about roots. Numerical methods can be used to approximate roots, but it is difficult to begin without knowing a real root exists. The intermediate value theorem can be used to show that every polynomial of odd degree (with real coefficients) has at least one real root. This is the best that can be expected, since a general polynomial of even degree may have no real roots, and a polynomial of odd degree may have exactly one real root.

Let $p : \mathbf{R} \to \mathbf{R}$ be a polynomial function of degree $n \geq 1$. The top degree term in a non-constant polynomial "dominates" the sum of the other terms, in the following sense:

$$p(x) = \sum_{k=0}^{n} a_k \, x^k = x^n \left(\sum_{k=0}^{n} a_k \, x^{k-n} \right), \tag{5.1}$$

and the term in parentheses has limiting value a_n as $|x| \to \infty$ since every other term goes to 0 (note that the exponents $k - n$ are *non-positive*). Since $a_n \neq 0$ by assumption, the term in parentheses has the same sign

as a_n provided $|x|$ is sufficiently large. Speaking "asymptotically," a non-constant polynomial behaves like its highest-degree term.

What bearing does this have on existence of roots of polynomials? First, we may as well assume $a_n > 0$, since multiplying by -1 does not change the set of roots of p. If $a_n > 0$, then $p(x) > 0$ for sufficiently large x; in fact, equation (5.1) asserts that $p(x) \to +\infty$ as $x \to +\infty$. However, if $a_n > 0$, then the behavior of $p(x)$ for large negative x depends on whether the degree is even or odd. If n is odd, then $p(x) \to -\infty$ as $x \to -\infty$. In this event, there exist real numbers $x_1 \ll 0$ and $x_2 \gg 0$ with $p(x_1) < 0 < p(x_2)$. By the intermediate value theorem, there exists an $x_0 \in (x_1, x_2)$ with $p(x_0) = 0$. In other words, a polynomial of odd degree has at least one real root.

By contrast, if n is even, then $p(x) \to +\infty$ as $x \to -\infty$, so there is no guarantee that p changes sign; for large $|x|$, the sign of $p(x)$ is the same as the sign of a_n, and there is no way to conclude that the sign changes. This says the *argument* we used for odd-degree polynomials fails, but conceivably there might be a "better" proof. To settle the matter conclusively, observe that the polynomial $p(x) = 1 + x^2$ has even degree and has no real roots, because the field $\mathbf{R}$ is ordered, and an ordered field cannot contain a square root of -1. In summary, a polynomial of odd degree has at least one real root (and generally has no more, see Exercise 5.8), and a general polynomial of even degree has no real roots.

The Bisection Method

The intermediate value theorem is the basis of a numerical recipe—the *bisection method*—for finding roots of a continuous function. The method only works for functions (such as polynomials) whose values are easy to calculate. To see how the method works, we will use it to approximate $\sqrt{2}$. Consider the function $f(x) = x^2 - 2$ with domain $[1, 2]$. By direct calculation, we find that $f(1) = -1 < 0$ and $f(2) = 2 > 0$. The intermediate value theorem implies there is a real zero in the open interval $(1, 2)$. Now we bisect; the midpoint is $3/2$, and we find that $f(3/2) = 9/4 - 2 = 1/4 > 0$. Since $f(1) < 0 < f(3/2)$, the intermediate value theorem asserts that f has a zero in the open interval $(1, 3/2)$. Again we bisect; the midpoint is $5/4$, and $f(5/4) = 25/16 - 2 < 0$, so we conclude that f has a zero in the interval $(5/4, 3/2)$, or in decimal, $(1.25, 1.5)$. The search pattern should be clear; you should carry out another step yourself, to ensure you understand.

Here is the general set-up and procedure, in algorithmic form suitable for writing a computer program. Suppose f is a continuous function on $[a, b]$, and that $f(a)$ and $f(b)$ have opposite sign. By the intermediate value theorem, there is a root in the interval (a, b). Evaluate f at the midpoint $(a+b)/2$; if the value is zero, then stop. Otherwise there is a sign difference between function values on exactly one of the half intervals, indicating that there is a root in that half interval. Repeat until the desired accuracy is obtained. This is a simple algorithm, but the accuracy only doubles with each iteration, so it takes about three or four iterations to get each extra decimal of accuracy. Other algorithms can give far higher accuracies; recall that by Exercise 4.23, the sequence in Example 4.47 converges to $\sqrt{b}$ in such a way that the *number of decimals of accuracy* doubles at each iteration. With well-chosen starting values, six iterations of the latter scheme can give 31 decimals, an accuracy requiring about 100 bisections. Ten iterations would give almost 500 decimals of accuracy, requiring over 1600 bisections. However, the bisection method has its uses, even though for practical calculations there are often better methods.

Exercises

Exercise 5.1 Show that if $f : \mathbf{R} \to \mathbf{R}$ is continuous and ℓ-periodic, then f is uniformly continuous on $\mathbf{R}$. ◇

Exercise 5.2 We saw that the restriction of the signum function to $\mathbf{R}^\times$ is locally constant but not uniformly continuous. Prove that if $a > 0$, then the restriction to $\mathbf{R} \setminus [0, a]$ *is* uniformly continuous. Note the distinction between removing a point and removing an arbitrarily small interval. ◇

Exercise 5.3 Find a bounded, continuous function $f : (0, 1) \to \mathbf{R}$ such that f is *not* uniformly continuous.
Suggestion: Let g be a non-constant, continuous, periodic function, and set $f(x) = g(1/x)$. ◇

Exercise 5.4 Let $p : \mathbf{R} \to \mathbf{R}$ be a polynomial function of degree at least 2. Prove that p is not uniformly continuous. ◇

Exercise 5.5 Let $I \subset \mathbf{R}$ be an open interval (possibly unbounded), and let $f : I \to \mathbf{R}$ be bounded, continuous, and increasing. Prove that f is uniformly continuous. ◇

Exercise 5.6 Let $I \subset \mathbf{R}$ be an *interval*, and let $f : I \to \mathbf{R}$ be increasing. Prove that f^{-1} is continuous.
Hint: Draw a sketch, and ask what it means for f^{-1} to be continuous. ◇

Exercise 5.7 Give an example of a continuous, increasing function whose inverse is discontinuous.
Hint: By the previous exercise, the domain cannot be an interval! ◇

Exercise 5.8 Prove that the polynomial $p(x) = x^5 + x - 1$ has a unique real root, and use the bisection method (and a calculator!) to approximate this root to two decimal places. If your calculator is *programmable*, then you should find as many decimal places as your calculator allows. In any case, writing a flow chart for an algorithm to obtain successively better approximations is a good exercise. ◇

Exercise 5.9 Let $p : \mathbf{R} \to \mathbf{R}$ be a polynomial function that is monic and of even degree ≥ 2.

(a) Prove that $\lim(p, \pm\infty) = +\infty$.

(b) Prove that p has an absolute minimum on $\mathbf{R}$, namely, there exists a real number x_0 such that $p(x_0) \leq p(x)$ for all real x.

Note that you cannot immediately apply the extreme value theorem in part (b); however, by carefully leveraging the information from part (a), you can reduce the search for a minimum to a closed, bounded interval. ◇

Exercise 5.10 Define $f : (0, +\infty) \to \mathbf{R}$ by $f(x) = \dfrac{5x^7 - 2x^4 - x^2 + 1}{x^7 + 2x^5 + 1}$.

(a) Prove that there is an $x_0 \in (0, +\infty)$ with $f(x_0) = 1$.

(b) Prove that f has an absolute minimum.

Hint: You will come to grief if you try to do either part without tools from this chapter. ◇

Exercise 5.11 This exercise is concerned with the existence of fixed points for continuous functions.

(a) Let $f : [a, b] \to [a, b]$ be a continuous function. Show that there exists an $x_0 \in [a, b]$ with $f(x_0) = x_0$. In words, every mapping of a closed, bounded interval to itself has at least one fixed point.

Suggestion: Consider the function g defined by $g(x) = f(x) - x$.

(b) Does the same conclusion hold if f maps some open interval to itself? What if $f : X \to X$, but X is not an interval?

Justify your answers in part (b). ◇

Exercise 5.12 Suppose $f : X \to \mathbf{R}$ is uniformly continuous on X, and let $(x_n)_{n=0}^{\infty}$ be a Cauchy sequence in X. Prove that the image sequence $\big(f(x_n)\big)$ is Cauchy. ◇

Exercise 5.13 Suppose $f : X \to \mathbf{R}$ is merely continuous on X, and let $(x_n)_{n=0}^{\infty}$ be a Cauchy sequence in X. Can you deduce that the image sequence $\big(f(x_n)\big)$ is Cauchy?
Suggestion: Start by trying to extend the proof you found for the preceding exercise. If you cannot generalize the proof, attempt to discover the step that does not work, and use the information to seek a counterexample. The most difficult part of this problem is deciding whether there is a proof or a counterexample! ◇

Exercise 5.14 Let $f : (a, b) \to \mathbf{R}$ be uniformly continuous.

(a) Prove that there is a continuous extension of f to $[a, b]$, namely, there exists a continuous function $F : [a, b] \to \mathbf{R}$ such that $F|_{(a,b)} = f$.
Hint: Use Exercise 5.12 and Theorem 4.45 to define $F(a)$ and $F(b)$.

(b) Is the inverse true? ("If f is not uniformly continuous, then there does not exist a continuous extension.") Give a proof or counterexample.

(c) Prove that the extension found in part (a) is unique, i.e., if F_1 and F_2 are continuous extensions of f to $[a, b]$, then $F_1 = F_2$.

This is an instance of a general principle in analysis: A uniformly continuous function has a unique continuous extension to the set of limit points of its domain. ◇

Exercise 5.15 Let $I \subset \mathbf{R}$ be an interval. A function $f : I \to \mathbf{R}$ is *Lipschitz continuous* if there exists a real number M such that

$$|f(x) - f(y)| \leq M\,|x - y| \qquad \text{for all } x,\, y \in I.$$

(a) Prove that if f is Lipschitz continuous, then f is uniformly continuous on I.

(b) Prove that $\sqrt{\ } : [0, \infty) \to \mathbf{R}$ is uniformly continuous, but not Lipschitz continuous.
Hint: The trouble occurs near 0.

(c) Prove that f is Lipschitz continuous iff

$$f(x + h) = f(x) + O(h) \quad \text{near } h = 0 \text{ for all } x \in I.$$

Give a geometric interpretation of Lipschitz continuity. ◇

Exercise 5.16 Let cb be the Charlie Brown function, and let $f(x) = \text{cb}(1/x)$ for $x \in (0, 1)$.

(a) Sketch the graph of f. Prove that for every $\delta > 0$, f maps the interval $(0, \delta)$ *onto* the interval $[0, 1]$.

(b) Define $g : (0, 1) \to \mathbf{R}$ by $g(x) = (1/x)f(x)$. Prove that for every $\delta > 0$, g maps the interval $(0, \delta)$ *onto* the interval $[0, \infty)$.

(c) Find a continuous function $h : (0, 1] \to \mathbf{R}$ whose image is the open interval $(-1, 1)$. Does there exist a continuous function $H : [0, 1] \to \mathbf{R}$ such that $h = H|_{(0,1]}$?
Hint: Constructing h is not as easy as it looks, but is relevant to this exercise.

Look at Exercise 5.3 if you haven't already. ◇

Exercise 5.17 Does there exist a bijective function $f : [0, 1] \to (0, 1)$? Does there exist a *continuous* bijective function $g : [0, 1] \to (0, 1)$? As always, give a proof or counterexample, as appropriate. ◇

Chapter 6

What is Calculus?

Calculus is the mathematical study of rates of change. The name refers to the calculational procedures—differentials, infinitesimals, and integrals—not the theoretical underpinnings with which we are principally concerned. Calculus is naturally divided into two halves: *differentiation*—the study of rates of change, and *integration*—the study of total change. If f is a "suitable" function, it makes sense to ask "how rapidly $f(x)$ changes as x varies." A non-trivial part of making this idea precise is determining which functions are "suitable," and defining what is meant by rate of change at a single point. Conversely, if the rate of change of f is known at each point of some interval, one might wish to determine the total change in f over the interval. Intuitively, one wishes to "add up" the instantaneous rates of change to get the total change.

The operations of differentiation and integration have definitions that are motivated by simple ideas, but which hide a number of theoretical and practical complexities. The aim of this short chapter is to motivate the coming theory with some intuitive arguments.

6.1 Rates of Change

Consider an automobile trip along a straight highway. The position of the car, x, is a function of time t, say $x = X(t)$. For definiteness, we agree that the trip starts at $t = 0$, and that $X(0) = 0$. The value of X at time t is the odometer reading (because we zeroed the odometer when we set out). A graph of the odometer reading as a function of time is a mathematical idealization of the trip.

Suppose we want to describe our speed in mathematical terms, using only the odometer reading and a stopwatch. Speed is defined as the rate of change of position with respect to time, so we observe the odometer at two different times, t and $t + \Delta t$, and compute

$$\text{Average speed over } [t, t+\Delta t] = \frac{X(t+\Delta t) - X(t)}{\Delta t}. \tag{6.1}$$

Experimentally, we think of our measurement "starting at t and continuing for Δt". You may have performed this experiment: Some highways have "measured miles"—signs placed one mile apart. If you drive at constant speed (usually 40 or 60 miles per hour) and time how long it takes to pass the signs, you can calibrate your speedometer.

If Δt is large, then (6.1) may be fairly inaccurate if our speed is not constant. A pair of measurements produces only the average rate of change, and fluctuations on a time scale smaller than Δt tend to "average out". In order to resolve shorter time intervals, we make Δt smaller. If we imagine dt to represent an "infinitesimal" time interval, equation (6.1) becomes

$$\text{Speed at time } t = \frac{X(t+dt) - X(t)}{dt}.$$

Geometrically, we are focusing attention on the graph of X inside smaller and smaller rectangles, namely are "zooming in". To say the rate of change exists amounts to saying that zooming in makes the graph of X look more and more like a line. This discussion applies to any quantity that varies as a function of any other quantity.

For some real-world phenomena, such as velocities of cars or planets, populations of species in an ecosystem, voltages in an electric circuit, concentrations of chemicals in a test tube, or air pressure along the leading edge of an airplane wing, the rate of change of a quantity is "well-behaved" in the sense that as the increment of the input variable becomes smaller, the average rate of change has a limit. Mathematically, we say that functions modeling these quantities are *differentiable*; they possess well-defined, finite rates of change. Roughly, the graph of a differentiable function is composed of infinitely many infinitesimal line segments. (This assertion is an extreme example of the Goethe quote, but is true enough to be useful as a heuristic.)

Classical physics (especially mechanics and electromagnetism) is concerned almost entirely with differentiable functions. The speed of a bullet can be measured accurately by stroboscopically photographing

it at two closely separated times and measuring how far it traveled in the interim. The location of a planet can be calculated very accurately, even predicted months or years into the future, if its positions at a few times are known. One of the stunning early successes of calculus and statistical analysis occurred in the 1830s, when astronomers discovered the asteroid Ceres, then subsequently lost it in the glare of the sun. Based on measurements of its past location, C. F. Gauss predicted accurately where Ceres could be found several months after it was last sighted.

Limitations of the Differentiable Functions

It is now widely recognized that many phenomena are not well-modeled by differentiable functions; examples include stock prices, Internet traffic, motion of the earth along an earthquake fault, the shapes of mountains and clouds, and positions of individual molecules in a glass of water. These phenomena are still modeled by continuous functions, but shrinking the temporal or spatial scale does not yield more accurate measurements of rates of change. Instead, difference quotients vary in a complicated way as the increment grows smaller, yielding unstable numerical values of the of rate of change. Geometrically, zooming in on the graph does not "smooth out" the variations. There is a trade-off when measuring rates of change if the quantity in question has large, small-scale fluctuations: The increment must be large enough to smooth out "noise," but small enough not to average out the rate of change one wants to measure.

This discussion is slightly over-simplified. For example, the earth's (human) population varies chaotically on a time scale of minutes, but is quite regular on a scale of years. Similarly, individual automobile accidents or house fires are very difficult to predict, yet an insurance company can say with great accuracy how many accidents and fires will occur per month, and how much the total claims will be. Individual molecules of water move chaotically, but a glass of water looks smooth and uniform to our eyes. (The small-scale complication is revealed by carefully putting in a drop of food coloring. If water behaved like an idealized model, the color would immediately dissipate throughout the glass.) All mathematical models have a "characteristic scale", outside of which the model fails to work well. Phenomena of classical physics are remarkable in the range of temporal and spatial scales where they are applicable.

6.2 Total Change

Returning to our car trip, suppose the speed of the car, s, is known as a function of time, say $s = S(t)$. The graph of S represents the speedometer reading. The distance traveled in the infinitesimal time interval between t and $t + dt$ is $S(t)\,dt$, and the total distance traveled up to time t_0 is obtained by "adding up" these infinitesimal distances as time ranges from 0 to t_0.

The formal algebraic calculation is compelling: The instantaneous speed satisfies

$$S(t) = \frac{X(t+dt) - X(t)}{dt} = \frac{dX}{dt}, \quad \text{or } S(t)\,dt = X(t+dt) - X(t) = dX.$$

Again, dt is supposed to be "infinitesimal", greater than zero but smaller than every positive real number. When we "add up" these terms for all $t \in [0, t_0]$, we find, to our great satisfaction, that the increments of X constitute a formally telescoping sum, cf. Exercise 4.29, and the "sum" is $X(t_0) - X(0)$, the distance traveled by the car!

6.3 Notation and Infinitesimals

We must surely be on the track of some interesting mathematics. However, one should have at least nagging doubts about this argument, since dt *cannot* be regarded as a real number, as it violates the trichotomy property. A great deal of controversy arose over just this issue: "What *is* an infinitesimal dt?" The standard treatment of calculus sidesteps this issue by relegating dt to the role of a convenient bookkeeping device. However, it is worth emphasizing that careful use of dt as an algebraic quantity—as in the argument above—both leads to quick "proofs" of many basic theorems of calculus and illuminates the meaning of the statements of these theorems. The very term "calculus" refers to the calculational procedures for manipulating infinitesimals *correctly*—in a way that does not contradict ordinary algebra.

In the final analysis, *logical consistency*, not intuitive plausibility, is the primary criterion for judging a mathematical idea. One way to prove logical consistency is to define all objects under consideration, and to formulate all of one's assertions, in terms of axioms known (or assumed) to be consistent. We have already reduced the notion of continuity of functions to axioms for the real numbers, which in turn

were built on axioms for the rational numbers, which were built upon arithmetic of the natural numbers, which finally was defined in terms of sets. Consistency of set theory was assumed. To justify the beautiful "verification" that position can be recovered from speed (or generally, that the total change of a quantity can be recovered by adding up infinitesimal increments), we must do one of the following:

- Define infinitesimals in terms of real numbers, and verify that they satisfy appropriate form rules of manipulation, such as the ordered field axioms. This is the approach of *non-standard analysis*.
- Use the real number axioms to define certain expressions (such as quotients and appropriate infinite sums) in which infinitesimals appear, and verify that these expressions can be manipulated *as if they contained actual infinitesimals*, while never actually using an expression in which a "naked" infinitesimal appears. For instance, if quotients of infinitesimals are defined in terms of the real number system, we need to prove rules like

$$\frac{dy}{dx} + \frac{dz}{dx} = \frac{dy + dz}{dx} \quad \text{and} \quad \frac{dz}{dy}\frac{dy}{dx} = \frac{dz}{dx}.$$

The first approach is a surprisingly difficult technical task, that requires adding a new axiom to *set theory itself*. Further, one provably gains nothing, in the sense that every "non-standard" theorem can be proven by "standard" means. (That said, a good case can be made that non-standard analysis is more intuitive, so that one is likelier to *find* theorems by non-standard techniques than without them.)

The seemingly convoluted second procedure is in fact the path we follow, which accounts for the indirect flavor that arguments tend to have. Most of the technical tools we need—chiefly the concept of limits and the means of manipulating them—have already been developed, so at this stage the indirect path is definitely more economical. The notation of infinitesimals is called *Leibniz notation* after Gottfried Wilhelm von Leibniz. Every expression in Leibniz notation can be written in a less provocative (i.e., non-infinitesimal) form, called *Newtonian notation* after Sir Isaac Newton. For example, a Newtonian writes $S(t) = X'(t)$ for the speed of the automobile above, while a devotee of Leibniz notation would write $s = dx/dt$. Mathematicians tend to prefer Newtonian notation, which is less tempting to abuse,

while scientists tend to prefer Leibniz notation because it allows them to translate real problems into symbolic form and to solve them with relative ease. You should strive for fluency in both languages, since their strengths are complementary.

Plan of Action

The definitions of integration and differentiation can seem a little complicated, but are just formalizations of the ideas discussed above. Working directly with the definitions is often difficult, but there are theorems that facilitate calculation of derivatives and integrals for a large class of functions.

The next four chapters explore the main aspects of these ideas, with the following results:

- Infinitely many instantaneous rates of change can be added to get a definite numerical result that represents total change.
- Rates of change can be manipulated like ratios of infinitesimal quantities.
- The operations of taking the rate of change and finding the total change are essentially inverse to each other.

Integration and differentiation may also be viewed as operations that create new functions from known functions. We will construct logarithms, exponential functions, and circular trig functions using the operations of calculus. The inverse relationship between derivatives and integrals will allow us to determine properties of functions so defined.

Chapter 7

Integration

We begin the study of calculus proper with the operation of *integration*. The intuitive wish is to begin with a function f defined on a closed, bounded interval $[a, b]$, and to "add up the infinitesimal terms $f(t)\,dt$ for each $t \in [a, b]$". As it stands, this goal is meaningless, because we do not know what "dt" stands for, and we do not know how to add infinitely many quantities. The quantity we hope to capture has a nice geometric interpretation, however. First consider a *non-negative* function f, and imagine dividing the domain $[a, b]$ into a large (but finite) number of subintervals of equal length. Construct rectangles as in Figure 7.1. The sum of the areas of these rectangles is "approximately" the quantity we want to define.

The sum of the areas is an approximation exactly because each rectangle has positive (real) width, not "infinitesimal" width. To get a "better" approximation, we should make the width smaller, that is, subdivide the domain into a larger number of subintervals, see Figure 7.2.

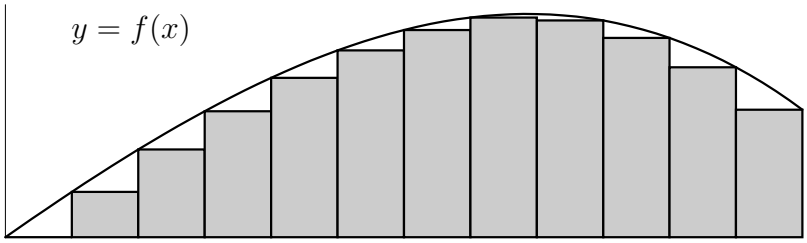


Figure 7.1: Areas of rectangles approximating the sum of $f(t)\,dt$ for $t \in [a, b]$.

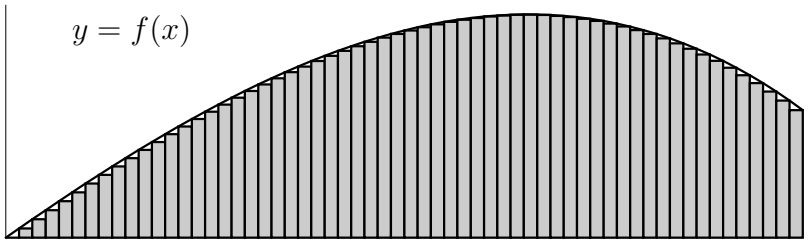


Figure 7.2: A larger number of rectangles gives a better approximation.

Of course, no matter how many rectangles we take, the resulting sum will probably not be exactly the quantity we wish to define. It is here that the completeness axiom for $\mathbf{R}$ saves the day; once things are set up properly, it will be easy to see that every one of our sums is smaller than a fixed real number. By the completeness axiom, the set of sums has a real supremum. If we are lucky, this supremum will be exactly the desired quantity!

This is another good time to re-consider the Goethe quote. Taking narrower rectangles gives a better approximation, so "in the limit" (taking the supremum) we are "adding up the areas of infinitely many infinitely thin rectangles". Observe that if we can make these ideas work, we will have (in a sense) assigned a real number to an expression of the form $0 \cdot \infty$.

It should be clear that the process just outlined has something to do with "the area under the graph of f". This is partly accidental, because we assumed that f was non-negative, and because Figures 7.1 and 7.2 depict a situation where the ideas work. It happens that if f is continuous, the process just outlined, called "integration", works well in the sense that the supremum really does represent something like "area". Nonetheless, it is quite difficult to use the definition to *compute* specific integrals, as we shall see. Fortunately, the idea of "summing the areas of infinitely many infinitesimal rectangles" is saved by two remarkable facts:

- It is possible to deduce many properties of integration without being able to evaluate a single integral.

- Using abstract properties of integration, we find that integration is closely related to the operation of "differentiation," which is more amenable to calculation.

These items are the subject of this and the following chapters, at the end of which we will have a powerful set of tools for describing and solving a wide variety of problems involving rates of change.

You should keep in mind that this chapter contains essentially no computational results. It is through a long chain of judicious definitions and hindsight observations that integration is at last forged into a useful computational tool. Put aside objections of impracticality for now. Integration is not conceptually difficult, but can seem daunting if you worry about how the definition will be used in practice. Be assured that you will soon have calculational techniques of great power and flexibility.

7.1 Partitions and Sums

Integration takes as input a bounded function $f : [a, b] \to \mathbf{R}$ and returns a real number. As described in the preceding section, we will aim indirectly for this target by precisely defining upper and lower bounds of the quantity we wish to define, then declaring that the quantity exists exactly when these bounds coincide.

Let $I = [a, b]$ be a closed, bounded interval. A *partition* of I is a finite collection of points $P = \{t_i\}_{i=0}^n$ such that

$$a = t_0 < t_1 < t_2 < \cdots < t_n = b.$$

The interval $I_i = [t_{i-1}, t_i]$ is called the ith *subinterval* of the partition, and has length

$$\Delta t_i = t_i - t_{i-1}.$$

We do not assume all the subintervals have the same length. The *mesh* of the partition P is the largest Δt_i, the length of the longest subinterval. The mesh of P is denoted $\text{mesh}(P)$ or $\|P\|$. If P and Q are partitions of I, and if $P \subset Q$, then Q is said to be a *refinement* of P. In other words, a refinement of P is obtained by adding finitely many points to P. Note that if $P \subset Q$, then $\text{mesh}(Q) \leq \text{mesh}(P)$; you cannot increase the mesh by adding points!

Now let $f : I \to \mathbf{R}$ be a *bounded* function. (If f is continuous on I, then f is bounded, by the extreme value theorem; however, at this stage the function f may be discontinuous everywhere.) Given a partition P of I, we take the "best" upper and lower bounds of f on

each subinterval:

$$\begin{aligned} m_i &:= \inf\{f(t) \mid t \in I_i\} \\ M_i &:= \sup\{f(t) \mid t \in I_i\} \end{aligned} \tag{7.1}$$

for $i = 1, \ldots, n$. Intuitively, m_i is the "minimum" of f on the ith subinterval, and M_i is the "maximum," but while f may have neither minimum nor maximum on I_i, we know the inf and sup exist because f is bounded. Using these lower and upper bounds of f, we form the *lower sum* and *upper sum* of f over the partition P by

$$L(f,P) = \sum_{i=1}^{n} m_i\,\Delta t_i, \qquad U(f,P) = \sum_{i=1}^{n} M_i\,\Delta t_i. \tag{7.2}$$

In the introduction, we considered only lower sums; for technical reasons that will shortly be apparent, we must consider both upper and lower bounds.

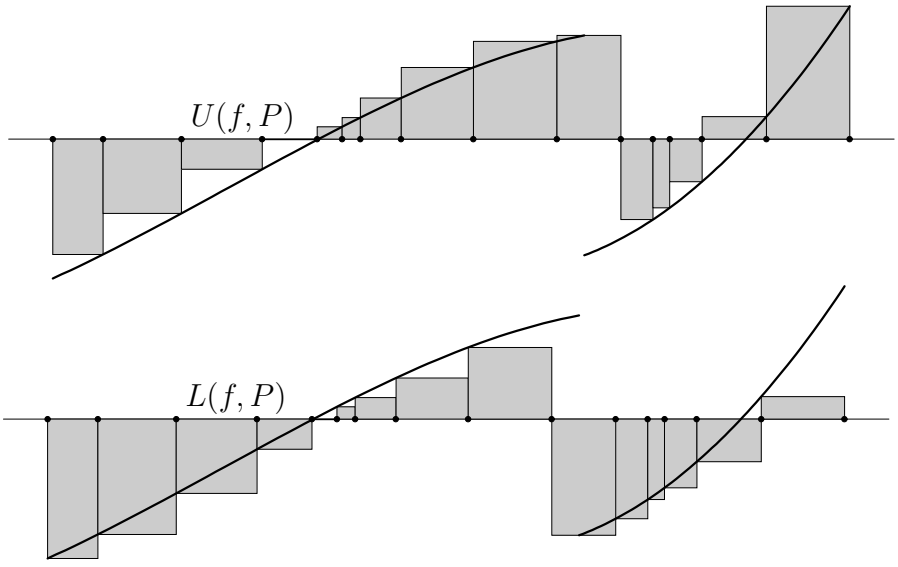


Figure 7.3: Upper (top) and lower sums of f associated to a partition.

Because $m_i \le M_i$ for each subinterval of P, it is clear that $L(f,P) \le U(f,P)$ for every f and every P. Further, for each i, $m_i \le f(t) \le M_i$ for all $t \in I_i$, so $m_i\Delta t_i$ is a reasonable lower bound for "the sum of $f(t)\,dt$ over $t \in I_i$," and similarly $M_i\Delta t_i$ is a reasonable upper bound.

The upper and lower sums may be regarded as the sums of areas of rectangles *provided f is non-negative*. Generally, a lower sum is the sum of the areas of the rectangles above the horizontal axis minus the sum of the areas below the axis, and similarly for an upper sum, see Figure 7.3. Refining the partition P can only improve the bounds. We formalize this useful observation as follows:

Lemma 7.1. *Let $f : [a, b] \to \mathbf{R}$ be a bounded function, and let P and Q be partitions of $[a, b]$ with $P \subset Q$. Then*

$$L(f, P) \leq L(f, Q) \leq U(f, Q) \leq U(f, P).$$

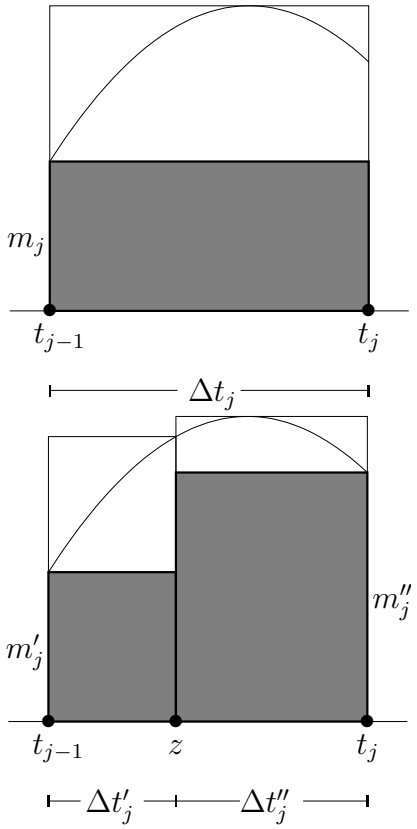


Figure 7.4: Refining the partition—before and after.

Proof. Since every refinement results from appending a finite number of points to P, induction on the number of additional points in Q

reduces the claim to the case where $Q = P \cup \{z\}$ has exactly one more point than P. Assume $z \in I_j$ for definiteness; the subdivision $I_j = [t_{j-1}, z] \cup [z, t_j]$ splits the term $m_j \Delta t_j$ in $L(f, P)$ into

$$m'_j \Delta t'_j + m''_j \Delta t''_j, \tag{7.3}$$

see Figure 7.4. But m'_j is the infimum of f on $[t_{j-1}, z] \subset I_j$, which is surely no smaller than m_j, and similarly $m''_j \geq m_j$. Consequently, the sum of the two terms in (7.3) is greater than of equal to $m_j \Delta t_j$. This is geometrically clear from Figure 7.4.

Since otherwise $L(f, P)$ and $L(f, Q)$ are identical, this argument proves that $L(f, P) \leq L(f, Q)$. A completely analogous argument shows that $U(f, Q) \leq U(f, P)$. □

Proposition 7.2. *Let $f : I \to \mathbf{R}$ be a bounded function, and let P and P' be* arbitrary *partitions of I. Then*

$$L(f, P) \leq U(f, P').$$

In words, every lower sum is less than or equal to every upper sum.

Proof. The partition $Q = P \cup P'$ is a refinement of both P and P'. By Lemma 7.1, $L(f, P) \leq L(f, Q)$ and $U(f, Q) \leq U(f, P')$. □

Given a bounded function $f : I \to \mathbf{R}$ whose domain is a closed, bounded interval, we have associated a set of lower sums (taken over all partitions) and a set of upper sums. Proposition 7.2 says that the set of lower sums is bounded above; any particular upper sum is an upper bound! Consequently, the set of lower sums has a real supremum $\mathbf{L}(f, I)$, which is called the *lower integral* of f on I. Dually, the set of upper sums is bounded below, hence has an infimum $\mathbf{U}(f, I)$, which is called the *upper integral* of f on I. Proposition 7.2 implies that $\mathbf{L}(f, I) \leq \mathbf{U}(f, I)$, which fortunately accords with our intuition: The quantity we are trying to define—the integral of f—should surely be no smaller than the lower integral and no larger than the upper integral.

Definition 7.3 Let I be a closed, bounded interval. A bounded function $f : I \to \mathbf{R}$ is *integrable* on I if $\mathbf{L}(f, I) = \mathbf{U}(f, I)$. In this case, the common value is called the *integral* of f over I.

The integral of f over I is denoted $\int_I f$, or by $\int_a^b f$ if $I = [a, b]$. It might seem intuitively obvious that the lower and upper integrals are equal, and though the proof is not obvious, the lower and upper

integrals do indeed coincide when f is *continuous*. As we will see by example, however, the lower integral of f is generally *strictly smaller* than the upper integral. In this event, the lower and upper integrals do not specify a unique real number, and we say that f is *not integrable* on I.

The definition of integrability relies on equality of the supremum of one set of numbers with the infimum of another set of numbers. For proving theorems, it is usually more convenient to use the following criterion, which uses one partition rather than on the set of all partitions.

Proposition 7.4. *A bounded function $f : [a,b] \to \mathbf{R}$ is integrable on $[a,b]$ if and only if the following condition holds:*
For every $\varepsilon > 0$, there exists a partition Q of $[a,b]$ such that

$$U(f,Q) - L(f,Q) < \varepsilon.$$

Proof. Suppose f is integrable. Fix $\varepsilon > 0$, and choose partitions P and P' such that

$$\left(\int_a^b f\right) - L(f,P) < \frac{\varepsilon}{2}, \qquad U(f,P') - \left(\int_a^b f\right) < \frac{\varepsilon}{2}.$$

Such partitions exist by definition of supremum and infimum. As in the proof of Proposition 7.2, take $Q = P \cup P'$, and conclude that $U(f,Q) - L(f,Q) < \varepsilon$.

Inversely, suppose f is *not* integrable. Let $2\varepsilon = \mathbf{U}(f,I) - \mathbf{L}(f,I)$. Then $U(f,Q) - L(f,Q) > \varepsilon > 0$ for every partition Q. □

7.2 Basic Examples

The integral of a function f depends only on the interval of integration; it is therefore sensible to write $\int_I f$. However, specific functions are usually given by formulas, like $f(t) = t^2$, and it would be convenient to write $\int_I t^2$. The problem is that the expression "t^2" does not define a function unless we agree that t^2 is the value of f *at* t, and adherence to such a convention is too much to ask, as will become apparent. What is needed is a "placeholder" to signify that t is the "variable" in the integrand. The standard notation is to write "$\int_I t^2\,dt$" in such a situation, the "dt" signifying that t^2 is the value of the integrand at t. This peculiar choice of notation is discussed at greater length below, but if you are literal-minded the interpretation here is sufficient. In

the expression $\int_I t^2\,dt$, t is a dummy variable, and (just as for limits) may be replaced by any other convenient symbol without changing the meaning of the expression.

It is instructive to see how the definition of integrability works by itself. Examples are given here to illustrate that the definition captures the notion of area in a couple of simple cases, and to show how a function can fail to be integrable.

Example 7.5 Let $c \in \mathbf{R}$, and let $C : [a, b] \to \mathbf{R}$ denote the corresponding constant function. For every partition of $[a, b]$ and for every subinterval, $m_i = c = M_i$. Consequently, every lower sum and every upper sum is equal to $c(b - a)$, so

$$\int_a^b C = \int_a^b c\,dt = c(b - a).$$

Observe that when $c > 0$, the value of the integral is the area of the rectangle enclosed by the t-axis, the graph of C, and the lines $t = a$ and $t = b$. When $c < 0$, the integral is minus the area of the rectangle.

The integral of the identity function f is comparatively painful to compute from the definition. However, elementary geometry suggests what the answer should be, so we have a sanity check for our result.

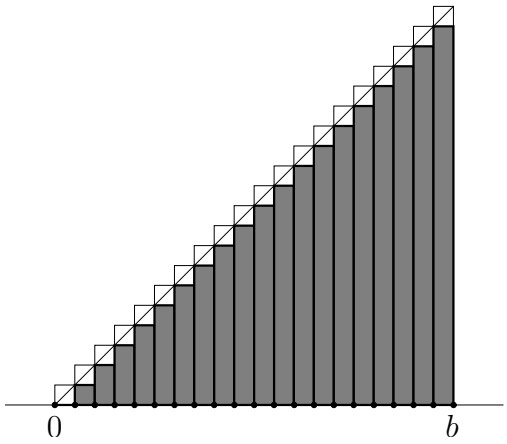


Figure 7.5: Lower and upper sums with $n = 20$ for the identity function.

Consider for simplicity an interval of the form $[0, b]$, and let $P_n = \{t_i\}$ be the partition with $(n + 1)$ equally-spaced points: $t_i = ib/n$, and $\Delta t_i = b/n$ for all i. The infimum and supremum of f on $[t_{i-1}, t_i]$

are t_{i-1} and t_i, respectively, so the lower and upper sums are (note the limits of summation)

$$L(f, P_n) = \sum_{i=1}^{n} t_{i-1}\,\Delta t_i = \sum_{i=0}^{n-1} \frac{ib}{n}\,\frac{b}{n} = \frac{b^2}{n^2}\sum_{i=0}^{n-1} i,$$

$$U(f, P_n) = \sum_{i=1}^{n} t_i\,\Delta t_i = \sum_{i=1}^{n} \frac{ib}{n}\,\frac{b}{n} = \frac{b^2}{n^2}\sum_{i=1}^{n} i.$$

These sums were evaluated in Exercise 2.6:

$$\sum_{i=1}^{n} i = \frac{(n+1)n}{2}, \qquad \sum_{i=0}^{n-1} i = \frac{n(n-1)}{2},$$

so

$$L(f, P_n) = \frac{b^2}{2}\left(\frac{n-1}{n}\right), \qquad U(f, P_n) = \frac{b^2}{2}\left(\frac{n+1}{n}\right).$$

From the first of these formulas, it is apparent that the supremum of the lower sums (over this particular family of partitions) is equal to $b^2/2$; the actual lower integral must therefore be at least $b^2/2$. Similarly, from the second formula it follows that the infimum of the upper sums (again taken over this family of partitions) is $b^2/2$, so the actual upper integral is no larger than $b^2/2$. In symbols,

$$\frac{b^2}{2} \le \mathbf{L}(f, I) \le \mathbf{U}(f, I) \le \frac{b^2}{2}.$$

But this means that the identity function is integrable, and that the integral over $[0, b]$ is equal to $b^2/2$, as expected from the geometric interpretation of the integral. □

Example 7.6 Let $f = \chi_{\mathbf{Q}}$ be the characteristic function of $\mathbf{Q}$. Then f is *not* integrable on the interval $[a, b]$, no matter how a and $b > a$ are chosen. Indeed, let P be a partition of $[a, b]$. In each subinterval, there exist rational numbers and irrational numbers, so f takes on the values 0 and 1 in *every* subinterval. But this means that $m_i = 0$ and $M_i = 1$ for every i, so

$$L(f, P) = \sum_{i=1}^{n} 0\,\Delta t_i = 0, \qquad U(f, P) = \sum_{i=1}^{n} 1\,\Delta t_i = (b - a),$$

regardless of P. The lower integral is therefore 0, while the upper integral is $b - a > 0$. Since these are unequal, f is not integrable. □

Example 7.7 One final but substantial example will illustrate how areas under curves were calculated before Newton. This serves the dual purpose of giving us a library of examples, and of emphasizing how difficult the definition is to use directly. (That said, the value of the result justifies the expense of effort.) Assume $0 < a < b$, and let k be a positive integer. Consider the monomial function $f(t) = t^k$ on $[a, b]$. We wish to calculate the integral $\int_a^b t^k \, dt$. Rather than use an "arithmetic" partition where all subintervals have equal length, we use a "geometric" partition where the ratio of the lengths of consecutive intervals is the same, Figure 7.6. The rationale is that the areas of consecutive rectangles will be in geometric progression because the integrand is a power function, so we can use the finite geometric sum formula to compute the lower and upper sums.

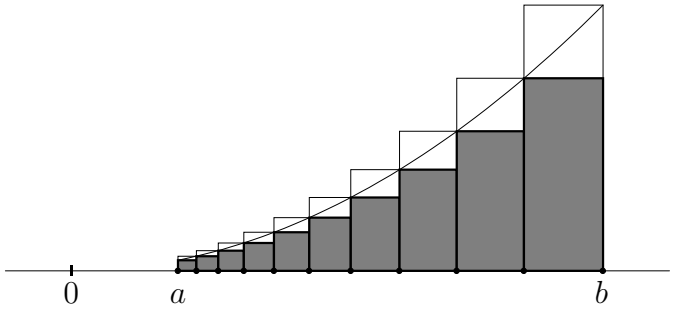


Figure 7.6: The lower and upper sums associated to a geometric partition.

Let $n > 0$ be the number of subintervals, and put $\rho = \sqrt[n]{b/a} > 1$, so that $b = a\rho^n$. The partition is $P = \{t_i\}_{i=0}^n$, with $t_i = a\rho^i$. The integrand is increasing, so the extrema on the interval $[t_{i-1}, t_i]$ are achieved at the endpoints:

$$m_i = (a\rho^{i-1})^k = a^k(\rho^k)^{i-1}, \quad M_i = (a\rho^i)^k \qquad \text{for } i = 1, \ldots, n.$$

Because $M_i = \rho^k m_i$ for all i, the upper sum is ρ^k times the lower sum. We will compute the upper sum, which is formally a little simpler. The general term is

$$f(t_i)\,\Delta t_i = (a\rho^i)^k \, a(\rho - 1)\rho^{i-1} = a^{k+1}\,\frac{\rho - 1}{\rho}(\rho^{k+1})^i,$$

so the geometric sum formula $\sum_{i=1}^{n} r^i = r\,\dfrac{r^n-1}{r-1}$ implies

$$U(f,P) = \sum_{i=1}^{n} f(t_i)\Delta t_i = a^{k+1}\,\frac{\rho-1}{\rho}\cdot\rho^{k+1}\cdot\frac{(\rho^{k+1})^n-1}{\rho^{k+1}-1}$$

Since $\rho^n = b/a$, we have $a^{k+1}\big((\rho^{k+1})^n-1\big) = b^{k+1}-a^{k+1}$, so that

$$U(f,P) = \rho^k\,(b^{k+1}-a^{k+1})\frac{\rho-1}{\rho^{k+1}-1}.$$

The reciprocal of the fraction is itself a geometric sum, $S(\rho) := \sum_{i=0}^{k}\rho^i$. Now, as $n\to+\infty$, the ratio $\rho = (b/a)^{1/n}$ approaches 1. Because S is a polynomial in ρ, we have $\lim(S,1) = \sum_{i=0}^{k} 1^i = k+1$. Consequently,

$$\mathbf{U}(f,I) \le \lim_{\rho\to1}\rho^k\,(b^{k+1}-a^{k+1})\frac{\rho-1}{\rho^{k+1}-1} = \frac{b^{k+1}-a^{k+1}}{k+1}.$$

To prove that this number really is the integral, just recall that the upper sum is ρ^k times the lower sum. As $n\to+\infty$, the lower sums tend toward the same limit as the upper sums, so

$$\frac{b^{k+1}-a^{k+1}}{k+1} \le \mathbf{L}(f,I).$$

As before, this simultaneously proves that f is integrable, and evaluates the integral. We shall see shortly that integrability can be deduced much more easily; the hard work here is needed to evaluate the integral. □

Differences of the form $F(b)-F(a)$ arise sufficiently frequently to warrant special notation: $F(x)\big|_{x=a}^{b} := F(b)-F(a)$. In this notation, Example 7.7 shows that

$$\int_a^b t^k\,dt = \frac{t^{k+1}}{k+1}\bigg|_{t=a}^{b} \qquad \text{for } 0<a<b. \tag{7.4}$$

The notation $\int_I f(t)\,dt$ is chosen for the intuition it overlies, that an integral is a sum of infinitely many infinitesimal terms $f(t)\,dt$. The infinitesimal dt may be viewed as a "renormalizing" factor, weighted so that $\int_a^b dt = b-a$. The notation is so compelling that it takes on a life of its own, and leads to reasonable but difficult-to-answer questions like, "What *is* 'dt'?" In this book, the infinitesimal under an integral sign is a placeholder and mnemonic device, but nothing more.

7.3 Abstract Properties of the Integral

Integration over a fixed interval $[a, b]$ can be viewed as a real-valued function whose domain is the set of functions that are integrable on $[a, b]$. This "integration functional" has several features in common with finite sums, which is fortunate for our wish that integration correspond (at least intuitively) to summing infinitesimals. First, it is *linear* in the sense of Chapter 3, see Theorem 7.8 below. Second, the integral of a non-negative function is non-negative. Third, integration satisfies an analogue of the triangle inequality, see Theorem 7.14. Finally, integration is "translation invariant" in the sense of Theorem 7.15.

Theorem 7.8. *Let f and g be integrable functions on an interval I, and let c be a real number. Then $f + g$ and cf are integrable, and*

$$\int_I (f+g) = \int_I f + \int_I g, \qquad \int_I (cf) = c\int_I f.$$

Proof. Let $P = \{t_i\}_{i=0}^n$ be an arbitrary partition of I, and set

$$m_i' = \inf_{t\in I_i}\{f(t)\}, \quad m_i'' = \inf_{t\in I_i}\{g(t)\}, \quad m_i = \inf_{t\in I_i}\{(f+g)(t)\},$$

The infimum of $f + g$ on I_i is at least as large as the infimum of f plus the infimum of g, namely $m_i \geq m_i' + m_i''$, and with analogous notation, $M_i \leq M_i' + M_i''$. Adding up these inequalities,

$$\begin{aligned} L(f,P) + L(g,P) &\leq L(f+g,P) \\ &\leq U(f+g,P) \quad \leq \quad U(f,P) + U(g,P) \end{aligned}$$

for every partition P. Taking suprema or infima as appropriate shows that

$$\int_I f + \int_I g \leq \mathbf{L}(f+g, I) \leq \mathbf{U}(f+g, I) \leq \int_I f + \int_I g. \tag{7.5}$$

This shows simultaneously that $f + g$ is integrable on I, and that the integral has the stated value. The assertion $\int_I (cf) = c\int_I f$ is Exercise 7.5. □

Example 7.9 Integrals, like limits, cannot distinguish functions that are equal except at finitely many points. Precisely, if $f : I \to \mathbf{R}$ is integrable, and if $g : I \to \mathbf{R}$ is equal to f except at finitely many points, then g is integrable, and $\int_I g = \int_I f$.

To prove this, consider the function $h = g - f$, which is zero except at finitely many points. If we can show h is integrable and has integral equal to zero, then the claim will follow by Theorem 7.8 because $g = h + f$. Because h is zero except at finitely many points, we may write

$$h = \sum_{j=1}^{k} c_j \, \chi_{\{x_j\}}$$

for some real constants c_j and distinct points x_j in I. It is therefore enough to show that each of the functions $\chi_{\{x_j\}}$ is integrable and has integral equal to zero. This is easy to do from the definition, see Exercise 7.6. □

The integral is *monotonic* in the sense that if f is a non-negative, integrable function on $[a, b]$, then $\int_a^b f \geq 0$. The following is a useful rephrasing; the proof is left to you (Exercise 7.7).

Theorem 7.10. *Let f and g be integrable functions on I. If $f(t) \leq g(t)$ for all $t \in I$, then $\int_I f \leq \int_I g$.*

In words, inequalities are preserved by integration over a fixed interval. One special case, a direct consequence of Theorem 7.10 and Example 7.5, is used repeatedly:

Corollary 7.11. *If $f : [a, b] \to \mathbf{R}$ is integrable, and if $m \leq f(t) \leq M$ for all $t \in [a, b]$, then*

$$m(b - a) \leq \int_a^b f \leq M(b - a).$$

A "patching property" of the integral is given in Theorem 7.12. The intuitive content is that to integrate a function over an interval we may split the interval into finitely many subintervals and sum the integrals over the separate pieces.

Theorem 7.12. *Let $f : [a, b] \to \mathbf{R}$ be a bounded function, and let $a < c < b$. Then f is integrable on $[a, b]$ if and only if f is integrable on both of the intervals $[a, c]$ and $[c, b]$, and in this case $\int_a^b f = \int_a^c f + \int_c^b f$.*

Proof. Suppose f is integrable on $[a, b]$. Fix $\varepsilon > 0$, and choose a partition P of $[a, b]$ such that $U(f, P) - L(f, P) < \varepsilon$. By adding the point c if necessary, we may assume $c \in P$. Let $P' \subset P$ be the set of points in $[a, c]$. From the definition it is clear that $U(f, P') - L(f, P') < \varepsilon$.

By Proposition 7.4, f is integrable on $[a, c]$. A similar argument shows f is integrable on $[c, b]$.

Conversely, suppose f is separately integrable on $[a, c]$ and $[c, b]$. Fix $\varepsilon > 0$ and choose respective partitions P' and P'' for $[a, c]$ and $[c, b]$ such that $U(f, P') - L(f, P') < \varepsilon/2$ and $U(f, P'') - L(f, P'') < \varepsilon/2$. The union $P = P' \cup P''$ is a partition of $[a, b]$ for which $U(f, P) - L(f, P) < \varepsilon$. This shows f is integrable on $[a, b]$ by Proposition 7.4.

In either case, $L(f, P) = L(f, P') + L(f, P'')$ and similarly for the upper sums, which proves $\int_a^b f = \int_a^c f + \int_c^b f$. □

Motivated by this result, we make the following *definitions* for an integrable function $f : I \to \mathbf{R}$:

$$\int_a^a f = 0, \quad \int_b^a f = -\int_a^b f \qquad \text{for all } a,\, b \in I. \tag{7.6}$$

With these definitions, the following "cocycle property" of the integral is easily checked.

Proposition 7.13. *Let $f : I \to \mathbf{R}$ be integrable, and let a, b, $c \in I$. Then*

$$\int_a^b f + \int_b^c f + \int_c^a f = 0.$$

The integral satisfies an analogue of the triangle inequality. As for finite sums, this result is a tool for estimating an integral in terms of the absolute value of the integrand.

Theorem 7.14. *If $f : I \to \mathbf{R}$ is integrable, then $|f| : I \to \mathbf{R}$ is integrable, and*

$$\left| \int_I f \right| \leq \int_I |f|.$$

Proof. The reverse triangle inequality says

$$\Big| |f(x)| - |f(y)| \Big| \leq |f(x) - f(y)| \quad \text{for all } x,\, y \in I. \tag{7.7}$$

Choose an arbitrary partition P of I, and let m_i and M_i be the infimum and supremum of f on the ith subinterval. Letting m_i' and M_i' denote the infimum and supremum of $|f|$ on the ith subinterval, equation (7.7) implies

$$M_i' - m_i' \leq M_i - m_i. \tag{7.8}$$

Now fix $\varepsilon > 0$ and choose a partition P such that $U(f,P) - L(f,P) < \varepsilon$. Equation (7.8) implies that for this partition, $U(|f|,P) - L(|f|,P) < \varepsilon$ as well. Since $\varepsilon > 0$ was arbitrary, $|f|$ is integrable.

The second part is now easy: $-|f(x)| \le f(x) \le |f(x)|$ for all x; by Corollary 7.11,

$$-\int_I |f| \le \int_I f \le \int_I |f|, \quad \text{or} \qquad \left|\int_I f\right| \le \int_I |f|,$$

as was to be shown. □

It is geometrically clear that if we "translate" the graph of f left or right, then integrate over appropriately shifted limits, the value of the integral is the same. This property is *translation invariance* of the integral.

Theorem 7.15. *If $f : [a,b] \to \mathbf{R}$ is integrable, and $c \in \mathbf{R}$, then*

$$\int_{a+c}^{b+c} f(s-c)\,ds = \int_a^b f(t)\,dt.$$

Proof. The letters s and t are used for variety, though they also suggest a "change of variable." The key observation is that if $P = \{t_i\}_{i=0}^n$ is a partition of $[a,b]$, then $P_c = \{t_i + c\}_{i=0}^n$ is a partition of $[a+c, b+c]$. If we set $g(s) = f(s-c)$, then clearly the infimum of g on $[t_{i-1}+c, t_i+c]$ is equal to the infimum of f on $[t_{i-1}, t_i]$ for all i, and similarly for the suprema. Consequently,

$$L(f,P) = L(g,P_c) \text{ and } P(f,P) = U(g,P_c)$$

for every P; the theorem follows immediately. □

Riemann Sums

Let $P = \{t_i\}_{i=0}^n$ be a partition of $[a,b]$, and let f be a bounded function on $[a,b]$. A *Riemann sum* taken from P is an expression of the form

$$\sum_{i=1}^n f(x_i)\,\Delta t_i, \quad x_i \in [t_{i-1}, t_i] \text{ for } i = 1, \ldots, n. \tag{7.9}$$

Since $m_i \le f(x_i) \le M_i$ for all i and all $x_i \in [t_{i-1}, t_i]$, every Riemann sum from P lies between $L(f,p)$ and $U(f,P)$. If one has bounds on

the lower and upper sums, then one can approximate an integral by any convenient Riemann sum. The practical advantage is that we can pick the x_i in any convenient way, and need know nothing about the inf or sup of f on the subintervals. Typical Riemann sums are given by $x_i = t_{i-1}$ (the *left-hand sum*), $x_i = t_i$ (the *right-hand sum*), and $x_i = (t_{i-1} + t_i)/2$ (the *midpoint sum*).

7.4 Integration and Continuity

This section contains two important technical results. The first, to the effect that continuous functions are integrable, gives a large class of integrable functions, though it does *not* directly give information on evaluating specific integrals. The second result asserts that a definite integral is a continuous function of its upper limit. The idea of regarding an integral as a function of its upper limit is fundamental, and is discussed at length below.

Theorem 7.16. *Let $f : [a, b] \to \mathbf{R}$ be a continuous function. Then f is integrable on $[a, b]$.*

Proof. We will show the upper and lower sums can be made arbitrarily close with a suitable choice of partition, thereby proving f is integrable by Proposition 7.4. The key fact is Theorem 5.5: A continuous function on a closed, bounded interval is *uniformly* continuous.

Fix $\varepsilon > 0$. By uniform continuity of f, there exists a $\delta > 0$ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \frac{\varepsilon}{2(b-a)}$. Choose an arbitrary partition P of mesh at most δ. For such a partition, the upper and lower sums are ε-close; indeed, if x and y are in the subinterval I_i, then $|x - y| < \delta$, so $|f(x) - f(y)| < \frac{\varepsilon}{2(b-a)}$. It follows that

$$M_i - m_i = \sup\{f(x) \mid x \in I_i\} - \inf\{f(x) \mid x \in I_i\} \leq \frac{\varepsilon}{2(b-a)}.$$

Since i was arbitrary, this inequality holds for all i, so we have

$$U(f, P) - L(f, P) = \sum_{i=1}^{n} (M_i - m_i)\Delta t_i \leq \frac{\varepsilon}{2(b-a)} \sum_{i=1}^{n} \Delta t_i = \frac{\varepsilon}{2} < \varepsilon.$$

By Proposition 7.4, f is integrable. □

Theorem 7.16 is a "hunting license" in the same way the extreme value theorem is: It asserts that certain functions are integrable, but does not say how to *find* the integral of a particular function. However, Theorem 7.16 does give a theoretical basis for *numerical approximation* of an integral, provided the integrand f is known explicitly. If the "winning strategy" of the proof can be implemented computationally, then the integral of f will be approximated to within ε by either an upper or lower sum for a convenient partition of mesh at most δ. The two corollaries below are restatements that are useful in applications. The first is often called *Riemann's theorem.*

Corollary 7.17. *Let $f : [a, b] \to \mathbf{R}$ be continuous. Fix $\varepsilon > 0$, and choose $\delta > 0$ such that*

$$|x - y| < \delta \implies |f(x) - f(y)| < \frac{\varepsilon}{2(b-a)}.$$

If P is a partition with $\|P\| < \delta$, then

$$\left| S - \int_a^b f \right| < \varepsilon \quad \textit{for every Riemann sum } S \textit{ taken from } P.$$

Corollary 7.18. *Let $f : [a, b] \to \mathbf{R}$ be continuous, and let (P_n) be a sequence of partitions—not necessarily nested—such that $\|P_n\| \to 0$ as $n \to \infty$. If S_n is a Riemann sum taken from P_n, then $\lim_{n\to\infty} S_n = \int_a^b f$.*

For example, $P = \{x_i\}_{i=0}^n$ might be the partition with equally-spaced points. There are much better numerical schemes for evaluating integrals, but they often work because the method can be proven to be better than the one given by the proof of Theorem 7.16.

The Integral as a Function of the Upper Limit

Let $f : [a, b] \to \mathbf{R}$ be integrable. For each $x \in [a, b]$, the function f is integrable on $[a, x]$ by Theorem 7.12. Define $F : [a, b] \to \mathbf{R}$ by

$$F(x) = \int_a^x f = \int_a^x f(t)\, dt. \tag{7.10}$$

A bit of thought confirms that F takes a single number as input, and returns a single number as output. Potentially, this process will produce new, interesting functions. However, the definition of F probably looks

strange; if x is given, how (in practical terms) is one to evaluate $F(x)$? This is our first serious example of a function that is not presented as an algebraic formula; we must appeal to the definition of the integral. To evaluate $F(x)$ from the definition for a single x, we must compute the supremum of the set of lower sums of f as P ranges over the set of partitions of $[a, x]$. As we have already seen in Examples 7.5 and 7.7, this is a relatively laborious and non-algorithmic task, even when f is a monomial. To see what we hope to gain, let us recall the result of Example 7.7:

$$\int_a^x t^k \, dt = \frac{x^{k+1} - a^{k+1}}{k+1} \qquad a,\ x > 0,\ k \in \mathbf{N}. \tag{7.11}$$

Depending how this equation is viewed, the result is either disappointing or intriguing. Perhaps, hoping to discover exotic new functions, we are disappointed to recover only a polynomial function as the integral of a monomial. We might, however, find it interesting that the integral on the left (a complicated object) is equal to the polynomial on the right (a simple object), and realize that this is a substantial and non-trivial piece of information. If you do not see why, it may be a good idea to review the philosophical points about functions made in Chapter 3. In particular, a single function may be described by two completely different "rules." The rule on the left-hand side of equation (7.11) is complicated, but has an interesting interpretation (the area of a certain non-polygonal region). The rule on the right is simplicity itself, but is of no particular intrinsic significance. That the two rules define the same function is truly useful! Suppose we wish to find the area of the region bounded by the t axis, the parabola $y = t^2$, and the lines $t = 1$ and $t = 2$. It is easy to express this area as an integral, but the integral is difficult to evaluate directly from the definition. In light of equation (7.11), we need not use the definition; we immediately read off that

$$\text{area} = \int_1^2 t^2 \, dt = \frac{2^3 - 1^3}{3} = \frac{7}{3}.$$

If we had the means to produce other "magic formulas" like this one, we would have a powerful computational tool at our disposal.

It turned out that the integral of the kth power function was not a new function, but in fact there are many simple integrals that *do* give rise to "exotic" functions that cannot be expressed through purely algebraic means. One of the most important non-algebraic functions is

the *natural logarithm*, defined by the innocuous-looking integral

$$L(x) = \int_1^x \frac{dt}{t}, \qquad t > 0.$$

In Exercise 7.17, you are asked to establish some basic properties of the natural logarithm. Aside from the details, you should note carefully how abstract properties of the integral are used to deduce facts about functions defined by integrals. Other interesting integrals are

$$\operatorname{asin}(x) = \int_0^x \frac{dt}{\sqrt{1-t^2}}, \qquad |x| < 1,$$

and

$$\operatorname{atan}(x) = \int_0^x \frac{dt}{1+t^2}, \qquad x \in \mathbf{R}.$$

Amazingly, these functions are related to circular trigonometry, but to see why, and to gain deeper understanding of functions defined as integrals, requires the material in Chapter 8.

Integration and O Notation

Our *proof* of equation (7.11) required the assumptions $0 < a \leq x$. In order to study how O notation behaves under integration, we need to extend our knowledge to the case $a = 0$. The proof tells us that both sides of (7.11) are continuous in a, so we may take the limit at $a = 0$ by evaluating.

Proposition 7.19. *If $k \geq 0$ is an integer and x is real, then*

$$\int_0^x t^k \, dt = \frac{x^{k+1}}{k+1}.$$

Proof. We first assume $x > 0$. Fix $a \in (0, x)$, and note that $0 \leq t^k \leq a^k$ for $t \in [0, a]$. Using the "trivial" partition $P' = \{0, a\}$, we have

$$0 = L(f, P') \leq U(f, P') = a^{k+1}.$$

Since refining improves the bounds, the same inequalities hold for *every* partition P' of $[0, a]$. Now let P'' be a partition of $[a, x]$; as in the proof of Theorem 7.12, if $P = P' \cup P''$, then

$$\begin{aligned} L(f, P'') &\leq L(f, P') + L(f, P'') = L(f, P) \leq U(f, P) \\ &= U(f, P') + U(f, P'') \leq a^{k+1} + U(f, P''). \end{aligned}$$

Taking the supremum of the lower sums and infimum of the upper sums (but keeping a fixed), we have

$$\frac{x^{k+1}-a^{k+1}}{k+1} \le L(f,P) \le U(f,P) \le a^{k+1} + \frac{x^{k+1}-a^{k+1}}{k+1}.$$

Letting $a \to 0$, we find that $\displaystyle\int_0^x t^k\,dt = \frac{x^{k+1}}{k+1}$ by the squeeze theorem.

The case $x < 0$ may be handled by repeating the calculation of Example 7.7, changing signs where appropriate. Alternatively, Proposition 7.13 and Exercise 7.13 imply that if $y > 0$, then

$$\int_0^{-y} t^k\,dt = -\int_{-y}^0 t^k\,dt = -\int_0^y (-t)^k\,dt = (-1)^{k+1}\int_0^y t^k\,dt,$$

which reduces the case $x < 0$ to the case $y > 0$. □

We next establish the fundamental principle that "integration increases the order of vanishing by one" in the following sense:

Theorem 7.20. *Let $k \in \mathbf{N}$, and let f be integrable on some interval containing a. If $f(x) = O\big((x-a)^k\big)$ on some interval I containing a, then $\int_a^x f = O\big((x-a)^{k+1}\big)$ on I.*

Proof. If we define $g(x-a) = f(x)$ then $g(u) = O(u^k)$, namely there exists a real number C such that $|g(u)| \le Cu^k$ for u in some interval about 0. Theorem 7.15 implies

$$\int_a^x f(t)\,dt = \int_0^{x-a} g(u)\,du,$$

so by the Theorem 7.14, Proposition 7.19, and equation (7.11) we have

$$\left|\int_a^x f(t)\,dt\right| = \left|\int_0^{x-a} g(u)\,du\right| \le C\cdot\left|\int_0^{x-a} u^k\,du\right|$$

$$\le \frac{C}{k+1}|u|^{k+1}\Big|_{u=0}^{x-a} = O\big((x-a)^{k+1}\big)$$

in some interval about a. □

Theorem 7.20 is useful, both theoretically and practically; it allows us to study integrals without dealing directly with εs and δs. To give a

simple but important application, we will prove that functions defined by integrals are automatically continuous. The proof foreshadows the so-called fundamental theorem of calculus, the key by which a large class of functions can be integrated.

Corollary 7.21. *Let $f : [a, b] \to \mathbf{R}$ be integrable. The function F defined by*

$$F(x) = \int_a^x f, \qquad x \in [a, b],$$

is continuous.

Proof. By assumption, f is bounded, that is, $f = O(1)$ on $[a, b]$. If x and $x + h$ are in $[a, b]$, then

$$|F(x + h) - F(x)| = \left| \int_a^{x+h} f - \int_a^x f \right| = \left| \int_x^{x+h} f \right| = O(h)$$

by the theorem. This proves not merely that F is continuous, but that F is Lipschitz (Exercise 5.15). Further, every bound on $|f|$ is a Lipschitz constant for F. □

You might wonder whether every continuous function is the integral of some other function. The answer is "no", because there exist continuous functions that are not Lipschitz. The square root function on $[0, 1]$ is an example.

An example will illustrate how F is found in concrete situations. Because we have relatively few calculational tools available, the example is (calculationally) extremely simple.

Example 7.22 Let $f : \mathbf{R} \to \mathbf{R}$ be the signum function, and let F be the integral, $F(x) = \int_0^x f$. Consider the cases $x > 0$ and $x < 0$ separately. Suppose $x > 0$. The signum function is equal to 1 on the half-open interval $(0, x]$, and is zero at 0. Since the integral is not altered by changing the value of f at finitely many points, we may as well assume f is equal to 1 on $[0, x]$. Thus

$$F(x) = \int_0^x f = \int_0^x 1 \, dt = x \quad \text{for } x \geq 0.$$

Similarly, if $x < 0$, then f is equal to -1 on the interval $[-1, x)$, and after changing the value at 0 we find that

$$F(x) = \int_0^x f = -\int_x^0 (-1) \, dt = -(-1)(0 - x) = -x \quad \text{for } x < 0.$$

In summary, $F(x) = |x|$ for $x \in \mathbf{R}$. □

7.5 Improper Integrals

Integration, by its very nature, requires a bounded function whose domain is a bounded interval. There are situations in which one wants to relax one or both of these requirements. *Improper integration* is a means of generalizing ordinary integrals to special situations where the integrand and/or interval of integration is unbounded. A few examples will illustrate the type of question we hope to answer.

The function $f : [0, 1] \to \mathbf{R}$ defined by

$$f(t) = \begin{cases} 1/t \text{ if } t > 0 \\ 0 \text{ if } t = 0 \end{cases}$$

is unbounded near 0, but locally bounded everywhere else. Suppose we wish to calculate the integral of f on $[0, 1]$. No matter what partition we pick, there is some subinterval on which f is unbounded, so it is impossible to compute an upper sum. A potential remedy is to take δ with $0 < \delta < 1$, regard the integral

$$F(\delta) := \int_\delta^1 \frac{1}{t}\, dt$$

as a function of δ, and to consider $\lim(F, 0^+)$. If this limit exists, then f is said to be "improperly integrable" on $[0, 1]$. By Exercise 7.17, the limit does not exist in this case, so the reciprocal function is not improperly integrable on $[0, 1]$. (The value of the integrand at 0 is immaterial; improper integrability is determined solely by "how rapidly" the integrand grows in absolute value near points where it is unbounded.)

If instead we wished to improperly integrate $f(t) = 1/\sqrt{|t|}$ over $[-1, 1]$, we would first split the interval of integration as $[-1, 0] \cup [0, 1]$ (to guarantee the integrand is only unbounded near one *endpoint* of each subinterval), then consider two separate improper subintegrals. If *both* improper subintegrals exist, then the original function is "improperly integrable." Unfortunately, we do not at this stage have the means to decide this question.

Our final example concerns the reciprocal function, but on the unbounded interval $[1, +\infty)$. The integrand is bounded, but it is impossible to partition the interval, because a partition has only finitely many points. The idea in this case is to attempt to define

$$\int_1^\infty \frac{1}{t}\, dt = \lim_{R \to +\infty} \int_1^R \frac{1}{t}\, dt.$$

Again by Exercise 7.17, the limit does not exist.

In general, an *improper integral* is an integral expression in which the integrand or interval of integration (or both) is unbounded. To decide whether an improper integral exists (or "converges"), the interval is split into finitely many subintervals such that on each piece either the interval is unbounded or the integrand is unbounded at exactly one endpoint, but not both. Each of the resulting improper integrals is considered separately. If *all* of them have a limit, then the sum of the limits is declared to be the value of the original expression. If one or more of the sub-problems fails to have a limit, then the original integral *does not exist*, or "diverges". It is not difficult to show that subdivision of the domain may be done in any convenient way, subject to the above criteria. The notation $\int_{\mathbf{R}}$ is sometimes encountered in lieu of $\int_{-\infty}^{+\infty}$.

Improper integrability is rarely decided by exact evaluation of the approximating "proper" integrals; rather, existence is deduced by appropriately *estimating* the approximating integrals. There is a useful *integral test* that relates summability of series and existence of improper integrals.

Proposition 7.23. *Let $f : [0, \infty) \to \mathbf{R}$ be a non-increasing, positive function, and let $a_k = f(k)$ for $k \in \mathbf{N}$. The sequence $(a_k)_{k=0}^{\infty}$ is summable iff f is improperly integrable. In other words,*

$$\sum_{k=0}^{\infty} a_k \text{ converges iff } \int_0^{\infty} f(t)\,dt \text{ converges.}$$

In Exercise 7.8 you will show that a non-increasing function is automatically integrable, so this hypothesis need not be made separately. The lower limit on the summation/integral is immaterial; convergence of an improper integral, like convergence of a series, is entirely contingent on the behavior of the tail, and has nothing to do with the behavior of the integrand on a bounded interval, or with finitely many terms of the series.

Proof. Let $k \in \mathbf{N}$. Because f is positive and non-increasing, we have

$$0 < a_{k+1} = f(k+1) \leq f(x) \leq f(k) = a_k \quad \text{for all } x \in [k, k+1].$$

Because the interval $[k, k+1]$ has length 1, the previous inequality integrates to

$$a_{k+1} \leq \int_k^{k+1} f(t)\,dt \leq a_k \quad \text{for all } k \in \mathbf{N}.$$

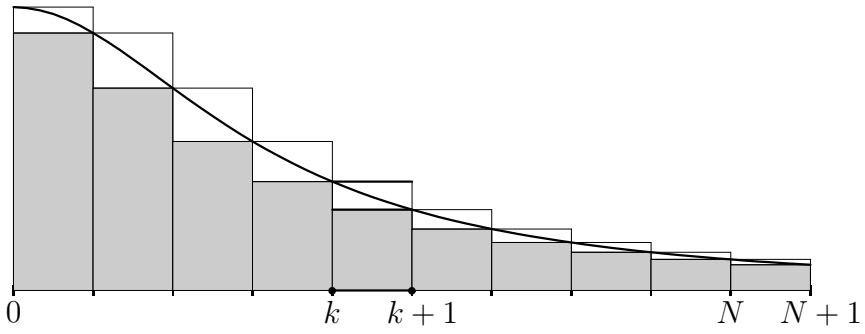


Figure 7.7: Bounding an integral above and below by partial sums of a series.

Summing these inequalities over $k = 0, \ldots, N$, we have

$$0 < \sum_{k=1}^{N+1} a_k = \sum_{k=0}^{N} a_{k+1} \leq \int_0^{N+1} f(t)\,dt \leq \sum_{k=0}^{N} a_k.$$

The series are lower and upper sums for the integral, see Figure 7.7. It follows that the integral is bounded as $N \to +\infty$ iff the partial sums of the series are bounded, which was to be shown. □

Using the "p-series test" (Example 4.64) and Proposition 7.23, you can easily determine when the improper integral

$$\int_1^{\infty} x^{-r}\,dx$$

converges, see Exercise 7.28. These integrals are very useful for estimating improper integrals with more complicated integrands. Improper integrals are a source of many delightful and ingenious formulas, but such applications must wait until we have a larger collection of functions at our disposal.

Exercises

Exercise 7.1 Let $f : [-2, 2] \to \mathbf{R}$ be the step function defined by

$$f(x) = \begin{cases} -1 & \text{if } -1 \leq x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Make a careful sketch of f, then sketch, on the same set of axes, the functions

$$F_0(x) = \int_0^x f, \qquad F_1 = \int_{-1}^x f.$$

Find an algebraic formula for F_1. ◇

Exercise 7.2 Let $f : [a, b] \to \mathbf{R}$ be a step function. Prove that the definite integral F is piecewise linear. ◇

Exercise 7.3 Modify the argument in Example 7.5 to evaluate

$$\int_0^x t^2\, dt \qquad \text{for } x > 0$$

directly. Give two general reasons that the squaring function is integrable on $[0, x]$. (The calculation of the area under a parabola is due to Archimedes of Syracuse.) ◇

Exercise 7.4 Suppose f is integrable on $[a, b]$, and that $c \in [a, b]$.

(a) Show that the functions defined by

$$F(x) = \int_a^x f(t)\, dt \quad \text{and} \quad G(x) = \int_c^x f(t)\, dt$$

differ by a constant. Hint: Use Proposition 7.13.

(b) If $H(x) = \int_x^b f(t)\, dt$, how are F and H related?

◇

Exercise 7.5 Complete the proof of Theorem 7.8 by showing that if f is integrable on I and $c \in \mathbf{R}$, then $\int_I (cf) = c \int_I f$.
Suggestion: The claim is obvious if $c = 0$. Consider the cases $c > 0$ and $c < 0$ separately. ◇

Exercise 7.6 Prove that the characteristic function of a point is integrable, with integral zero. More precisely, if $a \le x \le b$, then $\chi_{\{x\}} : [a, b] \to \mathbf{R}$ is integrable, and

$$\int_a^b \chi_{\{x\}} = 0.$$

◇

Exercise 7.7 Prove that if f is a non-negative integrable function on $[a, b]$, then $\int_a^b f \geq 0$. Use this result to prove Theorem 7.10. ◇

Exercise 7.8 Prove that a non-decreasing function $f : [a, b] \to \mathbf{R}$ is integrable.
Suggestion: Boundedness is clear. Consider partitions of $[a, b]$ into subintervals of equal length and write the difference of the lower and upper sums explicitly. ◇

Exercise 7.9

(a) Give an example of a non-decreasing function $f : [0, 1] \to \mathbf{R}$ that has infinitely many discontinuities.

(b) Prove that the denominator function of Example 3.11 is integrable on the interval $[0, 1]$. What is the value of the integral?

◇

Exercise 7.10

(a) Prove that a function $f : [a, b] \to \mathbf{R}$ is integrable if and only if the following condition holds: For every $\varepsilon > 0$, there exist step functions s_1 and s_2 such that $s_1 \leq f \leq s_2$ on $[a, b]$ and

$$\int_a^b (s_2 - s_1) < \varepsilon.$$

Intuitively, an integrable function can be "sandwiched" between step functions whose integrals are arbitrarily close.

(b) For each of the following functions f on $[0, 1]$, sketch the graphs of a pair of step functions as in part (a): The identity function; the characteristic function of the origin; the "$1/q$" function.

◇

Exercise 7.11

(a) Show that if f is integrable on $[a, b]$, then there exist *continuous* functions g and h such that $g \leq f \leq h$ on $[a, b]$ and

$$\int_a^b (h - g) < \varepsilon.$$

The result of the previous exercise and a sketch should be helpful.

(b)For each of the following functions f on $[0,1]$, sketch the graphs of a pair of continuous functions as in part (a): The identity function; the characteristic function of the origin; the "$1/q$" function.

◇

Exercise 7.12 Let f and g be integrable functions on $[a,b]$.

(a) Prove that f^2 is integrable on $[a,b]$.
Suggestion: By Theorem 7.14, $|f| \geq 0$ is integrable. Use $f^2 = |f|^2$ to bound the lower and upper sums.

(b) Prove that fg is integrable on $[a,b]$.
Hint: $2fg = (f+g)^2 - f^2 - g^2$.

The algebraic trick in part (b) is a "polarization identity". ◇

Exercise 7.13 Let a and b be real numbers, and let f be an integrable function on the closed interval with endpoints ca and cb for some real c. Prove that

$$\int_{ca}^{cb} f(t)\,dt = c\int_a^b f(ct)\,dt.$$

Suggestion: The case $c = 0$ is obvious. Successively consider the cases: $a < b$ and $c > 0$; $a < b$ and $c = -1$; $a < c$ and $c < 0$; $a > b$ and $c \in \mathbf{R}$. ◇

Exercise 7.14 Let $a > 0$, and let f be integrable on $[-a,a]$. Use Exercise 7.13 to prove the following:

(a) If f is odd, then $\displaystyle\int_{-a}^{a} f = 0$.

(b) If f is even, then $\displaystyle\int_{-a}^{a} f = 2\int_0^a f$.

(c) Let $F(x) = \int_0^x f$ for $x \in [-a,a]$. Prove that if f is even, then F is odd, and that if f is odd, then F is even.

This is merely an exercise is manipulating integrals; nothing technical is required. ◇

Exercise 7.15 Use Example 7.7 and Exercise 7.14 to show that

$$\int_0^x |t|\,dt = \frac{x|x|}{2} \qquad \text{for all } x \in \mathbf{R}.$$

Hint: Consider separately the cases $x \geq 0$ and $x < 0$. ◇

Exercise 7.16 Prove that

$$\int_a^b t^k \, dt = \frac{t^{k+1}}{k+1}\Big|_{t=a}^{b} \qquad \text{for all } a, b \in \mathbf{R}.$$

Suggestion: First use Example 7.7 and Proposition 7.19 to treat the case $a < 0$, $b = 0$; then split the integral at 0. ◇

Exercise 7.17 One of the most important functions of analysis is the *natural logarithm* function $\log : (0, \infty) \to \mathbf{R}$, defined[1] by

$$\log x = \int_1^x \frac{1}{t} \, dt.$$

(a) Use Proposition 7.13 to prove that log is an increasing function, and that $\log(1) = 0$.

(b) Prove that

$$\log(ab) = \log(a) + \log(b) \qquad \text{for } a, b > 0.$$

Suggestion: Write

$$\int_1^{ab} \frac{1}{t} \, dt = \int_1^a \frac{1}{t} \, dt + \int_a^{ab} \frac{1}{t} \, dt,$$

then use Exercise 7.13.

(c) Use part (b) to prove that $\log(1/a) = -\log a$ for all $a > 0$, and that more generally, $\log(a^n) = n \log a$ for all $a > 0$ and $n \in \mathbf{Z}$. Conclude that log maps $(0, \infty)$ *onto* $\mathbf{R}$.

(d) Prove that the reciprocal function is not improperly integrable on $[0, 1]$, nor on $[1, \infty)$.

(e) By part (c), there is a real number $e > 0$ with $\log(e) = 1$, and by part (a) this number is unique. Use explicit lower and upper sums to prove that $2 < e < 4$. A sketch will be immensely helpful.

[1] No mathematician calls this function ln outside a calculus course.

(f) In fact, $e < 3$; geometrically, the line tangent to the graph $y = 1/t$ at $t = 2$ lies below the graph, and encloses one unit of area between $t = 1$ and $t = 3$. Expressing this argument rigorously using only tools developed so far is not difficult. First prove that $1 - \frac{1}{2}t \leq \frac{1}{t}$ for all $t > 0$. Next, integrate both sides over $[1, 3]$, using Exercise 7.16 to handle the linear polynomial. To complete the proof, invoke an appropriate theorem from this chapter.

The number e plays a starring role in mathematics. In Chapter 12 we will find a series representation that converges rapidly. ◇

Exercise 7.18 By the proof of Proposition 7.23,

$$\sum_{k=2}^{n} \frac{1}{k} < \int_1^n \frac{1}{t}\, dt < \sum_{k=1}^{n-1} \frac{1}{k} \qquad \text{for all } n \geq 2.$$

(a) Make a careful sketch illustrating these inequalities.

(b) For $n \geq 2$, set

$$\gamma_n = \int_1^n \frac{1}{t}\, dt - \sum_{k=2}^{n} \frac{1}{k} = \log n - \sum_{k=2}^{n} \frac{1}{k}.$$

Prove that the sequence $(\gamma_n)_{n=2}^{\infty}$ is increasing and bounded above. Your sketch should suggest an inductive approach.

(c) By part (b), $\gamma := \lim_n \gamma_n$ exists. Determine whether or not γ is rational.[2]

The constant γ was introduced by Euler. ◇

The *average* of a finite list of numbers is the sum of the list divided by the number of entries. Analogously, the *average value* of an integrable function $f : [a, b] \to \mathbf{R}$ is defined by:

$$\text{Average value of } f \text{ on } [a,b] = \frac{1}{b-a}\int_a^b f = \int_a^b f \Big/ \int_a^b 1. \tag{7.12}$$

If f is non-negative, then the area under the graph is equal to the area of a rectangle of width $(b - a)$ and height equal to the average value of f. If the interval is fixed, the average may be denoted $\overline{f}$ or f_{avg}.

Exercise 7.19 If f is integrable on $[a, b]$, then $\int_a^b (f - \overline{f}) = 0$. ◇

[2]Resolving this open question will earn you an excellent publication and make you famous in mathematical circles.

Exercise 7.20 Let $f : [a, b] \to \mathbf{R}$ be continuous. For each positive integer n, let $P_n = \{t_i\}_{i=0}^n$ be the partition of $[a, b]$ into n intervals of equal length. Prove that

$$\lim_{n\to+\infty} \frac{1}{n+1} \sum_{i=0}^{n} f(t_i) = f_{\text{avg}}.$$

This further justifies the definition of "average value". ◇

Exercise 7.21 Prove that if $f : [a, b] \to \mathbf{R}$ is continuous, then there exists a $c \in (a, b)$ such that $f(c) = \overline{f}$. This result is called the *mean value theorem for integrals*.

If $f : [0, 1] \to \mathbf{R}$ is the squaring function, $f(x) = x^2$, find the value of c in $(0, 1)$ that satisfies the mean value theorem, and carefully sketch f and its average value. ◇

Exercise 7.22 Let f be integrable on some interval $(c - \eta, c + \eta)$.

(a) Prove that if $0 < |h| < \eta$, then

$$\frac{1}{h} \int_c^{c+h} f \tag{7.13}$$

is the average value of f on the closed interval with endpoints c and $c + h$ (even if $h < 0$).

(b) Assume f is continuous at c. Prove that

$$\lim_{h\to 0} \frac{1}{h} \int_c^{c+h} f = f(c).$$

(c) Show by example that the result of (b) fails in general if f is discontinuous at c.

◇

Exercise 7.23 Let $f : [a, b] \to \mathbf{R}$ be integrable. Prove that there exists an $x \in [a, b]$ such that

$$\int_a^x f(t)\, dt = \int_x^b f(t)\, dt.$$

Show by example that it is not generally possible to choose $x \in (a, b)$. ◇

Exercise 7.24 Let $f : [0, 1] \to \mathbf{R}$ be a function that is integrable on $[\delta, 1]$ for every δ in $(0, 1)$. Give a proof or counterexample to each of the following:

(a) If $\lim_{\delta \to 0} \int_\delta^1 f(x)\,dx$ exists, then f is integrable on $[0,1]$.

(b) If $\lim(f,0)$ exists, then f is integrable on $[0,1]$.

Note that f was not assumed to be bounded. ◇

Exercise 7.25 Let $g : [0,\infty) \to \mathbf{R}$ be non-negative and improperly integrable. Assume that $f : [0,\infty) \to \mathbf{R}$ is integrable on the interval $[0,x]$ for all $x > 0$, and that there exists an $R > 0$ such that $|f(t)| \le g(t)$ for $t \ge R$. Prove that f is improperly integrable on $[0,\infty)$. ◇

Exercise 7.26 Suppose f is integrable on $[0,x]$ for all $x > 0$. As usual, let

$$f_+ = \max(f,0), \qquad f_- = \min(f,0)$$

be the positive and negative parts of f. Prove that $|f|$ is improperly integrable on $[0,\infty)$ iff f_+ and f_- are improperly integrable on $[0,\infty)$. ◇

Exercise 7.27 Using Exercise 7.25 and part (d) of Exercise 7.17,

(a) Prove that $t \mapsto t^{-r}$ is *not* improperly integrable on $[1,\infty)$ for $r < 1$.
Hint: If $r < 1$ and $t \ge 1$, then $t^{-r} \ge t^{-1}$.

(b) Prove that $t \mapsto t^{-r}$ is *not* improperly integrable on $[0,1]$ for $r > 1$.

You do not need to know how to integrate $t^{-r}\,dt$, and should not use this knowledge if you have it. If you are fastidious, you may regard r as rational, since we have not yet defined t^r for irrational r. ◇

Exercise 7.28 Prove that if $r > 1$, then $\int_1^{+\infty} t^{-r}\,dt$ converges, subject to the same provisos as in the preceding problem. Use this result to show that

$$\int_0^{+\infty} \frac{dt}{1+t^2} \quad \text{and} \quad \int_2^{+\infty} \frac{dt}{t^2 - t}$$

converge. The first should be easy; the second is a little trickier, but not difficult if approached correctly. ◇

Chapter 8

Differentiation

Integration over an interval is a process of "putting together" infinitesimals $f(t)\,dt$ to obtain a real number. By varying the interval, we obtain a function. The other major operation of calculus is in a sense opposite. Differentiation is the process of finding the rate of change of a function, of "pulling apart" a function into infinitesimal increments. By varying the point at which the rate of change is taken, we obtain a new function that measures the rate of change of the original.

Newtonian and Leibniz Notation

For us, infinitesimals are a convenient fiction, and it is worth a short digression to re-discuss their status. The concept of "rate of change" is defined when a quantity y depends upon another quantity x, that is, when y is a function of x. Contemplation reveals that the central object of interest is the *function* itself, not the names we give to its input and output values. There are two notations prominent in calculus: Newtonian notation, in which the function is emphasized, and Leibniz notation, in which names of the inputs and outputs are emphasized. Each has merits and drawbacks:

- Newtonian notation is more compact and does not introduce the spurious symbol for the "independent variable", but does not suggest the infinitesimal nature of arguments.
- Leibniz notation is often easier to use for calculation and real-world modeling, but treats infinitesimals as if they were numbers, and assigns multiple meanings to symbols, leaving the user to read in the correct interpretation.

The less provocative Newtonian notation is analogous to the frame of a building. Its importance is usually not direct utility, but the way it unambiguously expresses the concepts of calculus in terms of numbers and functions, and the support it thereby gives to the friendlier but more easily abused Leibniz notation. Everyone uses Leibniz notation, but mathematicians unconsciously translate everything back to Newtonian language, especially when Leibniz notation falters. In order to use the calculus to full benefit, you should be fluent in both languages, and be able to translate freely between them. For that reason, the book develops the languages in parallel.

We have not defined infinitesimals,[1] and may therefore only use them for guidance, but not in definitions or proofs. Keeping this firmly in mind, it must be acknowledged that "calculus" in the traditional sense is precisely the manipulation of infinitesimal quantities. Several theorems that justify such manipulations are presented in this chapter and the next, and often the infinitesimal (Leibniz) interpretation is more compelling—and therefore easier to remember and use—than the limit-based (Newtonian) interpretation. For conceptual reasons alone (to say nothing of their calculational value), it is unwise to dispose of infinitesimals completely. In the final analysis, however, we must be certain that neither our definitions nor our arguments rest on anything but axioms for the real numbers. If infinitesimals are manipulated carelessly they lead to apparent paradoxes and other philosophical conundrums. In case of doubt, the definition is always the last word.

8.1 The Derivative

Suppose that "y is a function of x". The rate of change of y with respect to x is the ratio of the change in function values to the change in input values. Translating this into Newtonian language, if f is a function and if $[a, b]$ is an interval contained within the domain of f, then

$$\text{Average rate of change of } f \text{ over } [a,b] = \frac{f(b) - f(a)}{b - a}. \tag{8.1}$$

When the graph of f is a line (i.e., "the rate of change of f is constant"), the quotient above gives the slope of the line, in accord with intuition.

[1] More to the point, we have not shown that the existence of infinitesimals is logically consistent with the axioms of $\mathbf{R}$.

How should one define the rate of change of a function at a point? The naive answer, "Set $a = b$ in the formula above," is not helpful, for the right-hand side becomes the indeterminate expression $0/0$. In the early days of calculus, the "answer" was to take a and $b = a + dx$ to be infinitesimally close:

$$\text{Rate of change of } f \text{ at } a = \frac{f(a+dx) - f(a)}{dx} = \frac{dy}{dx}.$$

This idea works remarkably well in practice (if applied judiciously), but is subject to legitimate complaints. The symbol dx is meant to represent a positive quantity that is smaller than every positive real number. What *is* dx, then? If one's aim is to prove that calculus is free of logical contradiction, this objection is fatal. If the goal is simply to use calculus to describe the natural world, then the objection is moot as long as one's conclusions do not differ markedly from reality.

With the benefit of three centuries' hindsight (and the results of Chapter 4 at our disposal), we can neatly circumvent the objection. Let f be a function whose domain contains an interval $(x - \eta, x + \eta)$ for some $\eta > 0$. For the moment, the point x is arbitrary but fixed. If $0 < |h| < \eta$, then a *Newton quotient* of f at x is an expression

$$\frac{\Delta y}{\Delta x}(x, h) := \frac{f(x+h) - f(x)}{h}, \tag{8.2}$$

see Figure 8.1. The Newton quotient in (8.2) is the average rate of change of f on the interval with endpoints x and $x + h$, cf. (8.1). You should verify that this is true even when $h < 0$.

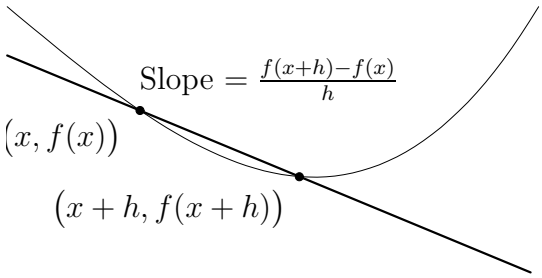


Figure 8.1: The Newton quotient as an average rate of change.

For simplicity, we may write $\Delta_x f(h)$ instead of $\frac{\Delta y}{\Delta x}(x, h)$.[2] For each x in the domain of f, $\Delta_x f$ is a function whose domain consists of all non-

[2]The notation $\Delta_x f$ is not standard in this context.

zero h such that $x + h$ is in the domain of f. By assumption, $\Delta_x f$ is defined on a deleted interval about 0, but is not defined at 0; however, $\Delta_x f$ may well have a *limit* at 0.[3] If

$$\lim(\Delta_x f, 0) = \lim_{h \to 0} \frac{\Delta y}{\Delta x}(x, h)$$

exists, then the limit is denoted $f'(x)$ or $\frac{dy}{dx}(x)$ in Newton and Leibniz notation respectively, and called the *derivative* of f at x, Figure 8.2. In this event, f is said to be *differentiable at* x, and the derivative is interpreted as the "instantaneous rate of change" of f at x.

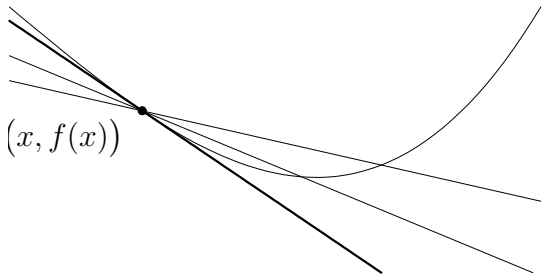


Figure 8.2: The tangent line as a limit of secant lines.

The Leibniz notation $\frac{dy}{dx}$ for the derivative suggests a quotient of infinitesimals, namely the infinitesimal increment of y divided by the corresponding infinitesimal increment in x. Though we do not define the symbols dy and dx individually, we *do* assign a precise meaning to their "quotient": The latter is the *limit* of ratios $\frac{\Delta y}{\Delta x}$ as $\Delta x \to 0$. We must be wary of using familiar-looking manipulations on these quotients, however. Before we may use identities such as

$$\frac{d(y+z)}{dx} = \frac{dy}{dx} + \frac{dz}{dx} \quad \text{or} \quad \frac{dz}{dy}\frac{dy}{dx} = \frac{dz}{dx},$$

we must revert to the definitions of these expressions as limits of real quotients to verify whether or not such equations are indeed true. At present we have no logical basis for assuming such formulas extend to infinitesimal quotients.

[3] In this assertion is resolved one of Zeno's paradoxes of motion, as well as the heated debate between Newton and Bishop Berkeley on the nature of infinitesimals.

Derivatives and o Notation

The "calculi of sloppiness" we introduced in Chapter 4 come into their own in differential calculus. In o notation, if f is differentiable at x, then

$$\frac{f(x+h)-f(x)}{h} = f'(x) + o(1) \quad \text{at } h = 0.$$

Multiplying by h and adding $f(x)$ to both sides, we find that if f is differentiable at x, with derivative $f'(x)$, then

$$f(x+h) = f(x) + h\, f'(x) + o(h) \quad \text{at } h = 0. \tag{8.3}$$

This argument can be run in the other direction as well; if f is defined on a neighborhood of x, and if $f(x+h) = f(x)+h\,c+o(h)$ at $h = 0$, then f is differentiable at x and $f'(x) = c$. Equation (8.3) is an extremely useful reformulation of the definition of differentiability: In order to prove a function ϕ is differentiable at x, we need only show that there exists a number c such that $\phi(x+h) = \phi(x) + h\,c + o(h)$ near $h = 0$. Further, if we can express c in terms of known quantities, then we have found $\phi'(x)$.

The definition of the derivative is brief (unlike the definition of the integral), but deceptively simple. Much of the technical machinery of Chapter 4 is involved, and there are deep consequences of the definition that seem intuitively plausible but require the results of Chapter 5. These deeper properties are collected in Chapter 9. This chapter is concerned with the elementary aspects of differentiation, which often turn out to be simple, calculational consequences of (8.3). Compared to integration, differentiation is relatively algorithmic from the definition. Derivatives of sums, products, quotients, and compositions of differentiable functions can be calculated with a few easily-memorized formulas. Differentiability at a point manifests itself geometrically as existence of a line "tangent to" the graph, and the sign of $f'(x)$ tells whether the function is increasing or decreasing at x in a sense.

Proposition 8.1. *If f is differentiable at x, then f is continuous at x.*

Proof. By assumption, the domain of f contains some interval about x, so x is a limit point of the domain, and it makes sense to ask whether or not $\lim(f, x) = f(x)$. But since f is differentiable at x, we have

$$f(x+h) = f(x) + h\, f'(x) + o(h) = f(x) + O(h) + o(h) = f(x) + O(h).$$

This implies f is continuous at x. □

The converse of Proposition 8.1 is false. The absolute value function, $f(x) = |x|$, is continuous at 0, but not differentiable at 0. Indeed, the Newton quotient is

$$\frac{f(h) - f(0)}{h - 0} = \frac{|h|}{h} = \operatorname{sgn} h,$$

the signum function, which has no limit at 0. Generally, a continuous function is differentiable *nowhere*, though we will not exhibit an example until Chapter 11.

Among the most basic functions are monomials, for which there is a simple differentiation formula:

Proposition 8.2. *Let $f(x) = x^n$ for n a positive integer. Then f is differentiable everywhere, and $f'(x) = nx^{n-1}$ for every $x \in \mathbf{R}$. In Leibniz notation, $\frac{d(x^n)}{dx} = nx^{n-1}$.*

In particular, the derivative of a monomial is a monomial of degree one lower. The formula extends to $n = 0$ if we agree that $0x^{-1} = 0$ for all x. We will see presently that this result allows us to differentiate polynomial functions with ease.

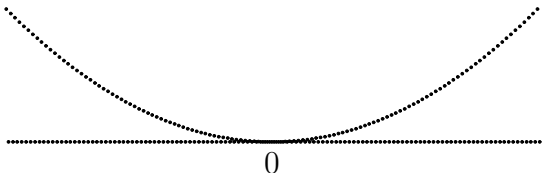
Proof. The binomial theorem implies

$$(x + h)^n = x^n + h\, nx^{n-1} + O(h^2) = x^n + h\, nx^{n-1} + o(h)$$

at $h = 0$. The proposition follows at once. □

Example 8.3 A function $f : \mathbf{R} \to \mathbf{R}$ can be differentiable at exactly one point, and discontinuous at every other point. An example is

$$f(x) = \begin{cases} x^2 \text{ if } x \in \mathbf{Q}, \\ 0 \text{ if } x \notin \mathbf{Q}. \end{cases}$$



If $x \neq 0$, then $\lim(f, x)$ does not exist by Corollary 4.21. Since f is not continuous at $x \neq 0$, *a fortiori* f is not differentiable at x, by Proposition 8.1. To differentiate f at 0, compute the Newton quotient:

$$\Delta_0 f(h) = \frac{f(h) - f(0)}{h} = \begin{cases} h \text{ if } h \in \mathbf{Q}, \\ 0 \text{ if } h \notin \mathbf{Q}. \end{cases}$$

Since $0 \le \Delta_0 f(h) \le |h|$ for all $h \neq 0$, the squeeze theorem implies $\lim(\Delta_0 f, 0)$ exists and is equal to 0; in other words, $f'(0) = 0$.

Generally, let $f : (-\eta, \eta) \to \mathbf{R}$ be a function such that $f(h) = O(h^2)$ near $h = 0$. Geometrically, the graph of f lies between a pair of parabolas of the form $y = \pm Cx^2$. Since $f(0)$ must be 0, we have $f(h) = f(0) + h\,0 + o(h)$, which shows that $f'(0)$ exists and is equal to 0. □

Derivatives of Sums, Products, and Quotients

As mentioned, there are calculational rules for differentiating a sum, product, or quotient of differentiable functions. As an easy translation exercise, you should express the conclusions of the following results in Leibniz notation.

Proposition 8.4. *Suppose f and g are differentiable at x, and that $c \in \mathbf{R}$. Then the functions $f + g$ and cf are differentiable at x, with derivatives given by $(f + g)'(x) = f'(x) + g'(x)$ and $(cf)'(x) = c\,f'(x)$.*

Proof. By hypothesis, there exist real numbers $f'(x)$ and $g'(x)$ such that

$$\left.\begin{aligned} f(x+h) &= f(x) + h\,f'(x) + o(h) \\ g(x+h) &= g(x) + h\,g'(x) + o(h) \end{aligned}\right\} \quad \text{at } h = 0. \tag{8.4}$$

Adding these equations, we have

$$(f+g)(x+h) = (f+g)(x) + h\,\big(f'(x) + g'(x)\big) + o(h).$$

This simultaneously proves that $f + g$ is differentiable, and that the derivative at x is $f'(x) + g'(x)$. The assertion for constant multiples is similar and left to you. □

Corollary 8.5. *If $p(x) = \sum_{k=0}^{n} a_k x^k$ is a polynomial, then p is differentiable at x for all $x \in \mathbf{R}$, and*

$$p'(x) = \sum_{k=1}^{n} k a_k x^{k-1}$$

For example,

$$\frac{d}{dx}(1 - x + x^2 - x^3 + x^4) = -1 + 2x - 3x^2 + 4x^3,$$
$$\frac{d}{dx}(x + 2x^2 + 3x^3) = 1 + 4x + 9x^2.$$

Let $X = (a, b)$ be an open interval, and let $\mathscr{D}^1(X) \subset \mathscr{F}(X, \mathbf{R})$ denote the set of differentiable functions on X. Proposition 8.4 says that

- $\mathscr{D}^1(X)$ is a vector subspace (Chapter 3), and
- The mapping $f \in \mathscr{D}^1(X) \mapsto Df := f' \in \mathscr{F}(X, \mathbf{R})$ is linear.

We will see presently that the derivative of a differentiable function is not generally differentiable, so D is technically not an operator on $\mathscr{D}^1(X)$. In fact, the image of D is too complicated to characterize in this book.

The Product and Quotient Formulas

There are analogous formulas for products and quotients that are easily calculated with o notation. It is convenient to establish a general reciprocal formula first.

Lemma 8.6. *If $f(x + h) = 1 + ah + o(h)$ near $h = 0$, then*

$$\frac{1}{f(x + h)} = 1 - ah + o(h)$$

near $h = 0$.

Proof. In some interval about $h = 0$, we have $|ah + o(h)| < 1$. The geometric series formula gives

$$\frac{1}{f(x + h)} = \frac{1}{1 + ah + o(h)}$$
$$= 1 - \big(ah + o(h)\big) + \big(ah + o(h)\big)^2 - \cdots = 1 - ah + o(h)$$

near $h = 0$. □

Theorem 8.7. *Suppose f and g are differentiable at x. Then fg is differentiable at x, and*

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x).$$

If $g(x) \neq 0$, then f/g is differentiable at x, and

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}.$$

Proof. Equation (8.4) holds by assumption, so

$$\begin{aligned}(fg)(x+h) &= \big[f(x) + h\,f'(x) + o(h)\big]\big[g(x) + h\,g'(x) + o(h)\big] \\ &= (fg)(x) + h\,\big[f'(x)g(x) + f(x)g'(x)\big] + o(h)\end{aligned}$$

which establishes the product rule. We break the argument for quotients into two steps, first treating simple reciprocals.

For brevity, write $a_0 = g(x) \neq 0$ and $a_1 = g'(x)$. Differentiability of g says $g(x+h) = a_0 + h\,a_1 + o(h) = a_0\big(1 + h\,(a_1/a_0) + o(h)\big)$ near $h = 0$. By Lemma 8.6,

$$\frac{1}{g(x+h)} = \frac{1}{a_0} \cdot \frac{1}{1 + h\,(a_1/a_0) + o(h)} = \frac{1}{a_0} - h\,\frac{a_1}{{a_0}^2} + o(h) \quad \text{near } h = 0.$$

It follows that $1/g$ is differentiable at x, and that

$$\left(\frac{1}{g}\right)'(x) = -\frac{a_1}{{a_0}^2} = -\frac{g'(x)}{g(x)^2}.$$

The general result now follows by writing $f/g = f \cdot (1/g)$ and using the results just proven:

$$\begin{aligned}\left(\frac{f}{g}\right)'(x) &= f'(x)\frac{1}{g(x)} + f(x)\left(\frac{1}{g}\right)'(x) \\ &= f'(x)\frac{1}{g(x)} - f(x)\frac{g'(x)}{g(x)^2} = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}\end{aligned}$$

as claimed. □

It is possible to differentiate polynomials such as $p(x) = x^{n+m} = x^n x^m$ or $q(x) = (1-x)(1+x^2)$ in two different ways, either by multiplying out and using Corollary 8.5, or with the product rule. You should verify that the two methods yield the same answer.

Theorem 8.7 allows us to extend the monomial differentiation formula to terms with negative exponent. The proof is left as an exercise.

Proposition 8.8. *If $f(x) = x^n$ for n an integer, then $f'(x) = nx^{n-1}$ for all $x \neq 0$.*

It follows from Theorem 8.7 and Corollary 8.5 that every rational function is differentiable in its natural domain. For example,

$$\frac{d}{dx}\frac{1}{1+x^2} = -\frac{2x}{(1+x^2)^2}, \qquad f(x) = 1,\ g(x) = 1+x^2,$$
$$\frac{d}{dx}\frac{x}{1+x^2} = \frac{(1+x^2)-x(2x)}{(1+x^2)^2} = \frac{1-x^2}{(1+x^2)^2}, \quad f(x) = x,\ g(x) = 1+x^2.$$

Note that algebraic manipulations may make a rational function easier to differentiate. For example,

$$\frac{x^2-1}{x^2+1} = 1 - \frac{2}{x^2+1},$$

but the right-hand side is easier to differentiate than the left-hand side.

Differentiating Integrals

In Chapter 7 we saw that integration can be used to construct new functions from known ones: If f is integrable on some interval $[a, b]$, then the equation

$$F(x) = \int_a^x f, \qquad x \in [a, b],$$

defines a continuous function on $[a, b]$. It is natural to attempt to differentiate F, and to expect a formula for F' in terms of f. If you have done Exercise 7.22, you already know the outcome.

Theorem 8.9. *Let $f : [a, b] \to \mathbf{R}$ be an integrable function, and suppose f is continuous at $c \in (a, b)$. Then the function F defined above is differentiable at c, and $F'(c) = f(c)$.*

This theorem may seem an amusing curiosity, but as we shall see it earns its name, the *fundamental theorem of calculus*. We are not yet in a position to understand its full significance, but certainly it indicates a close relationship between integration and differentiation. The proof was outlined in Exercise 7.22, but here are a few more details.

Proof. The cocycle property of the integral (Proposition 7.13) says that $\int_a^{c+h} = \int_a^c + \int_c^{c+h}$ as long as $c+h$ is in $[a,b]$. In other words,

$$F(c+h) - F(c) = \int_a^{c+h} f - \int_a^c f = \int_c^{c+h} f.$$

The Newton quotient for F at c is therefore given by

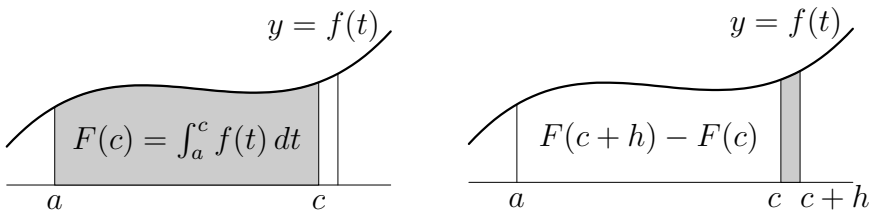


Figure 8.3: The increment of a definite integral.

$$\Delta_c F(h) = \frac{F(c+h) - F(c)}{h} = \frac{1}{h} \int_c^{c+h} f,$$

the average value of f on the interval with endpoints c and $c+h$. Now we use continuity: $f = f(c) + o(1)$ at c, so for h near 0 we have

$$\int_c^{c+h} f = \int_c^{c+h} \big(f(c) + o(1)\big) = h\, f(c) + o(h),$$

see Theorem 7.20. Therefore, $\Delta_c f(h) = f(c) + o(1)$ near $h = 0$. □

Nothing can be said if f is not continuous at c; examples show that F may or may not be differentiable at c. The signum function, with a jump discontinuity at 0, integrates to the absolute value function, which is not differentiable at 0. By contrast, if f is zero everywhere but the origin, and $f(0) = 1$, then the integral of f over an arbitrary interval is zero, so F is the zero function, which is clearly differentiable even at the origin.

The Chain Rule

The Chain Rule is a formula for the derivative of the composition of two differentiable functions.

Theorem 8.10. *Suppose f is differentiable at x and that g is differentiable at $f(x)$. Then $g \circ f$ is differentiable at x, and*

$$(g \circ f)'(x) = g'\big(f(x)\big) \cdot f'(x).$$

Proof. By hypothesis,

$$\begin{aligned} f(x+h) &= f(x) + h\, f'(x) + o(h) \quad \text{at } h = 0 \\ g(y+k) &= g(y) + k\, g'(y) + o(k) \quad \text{at } k = 0 \end{aligned}$$

If we write $y = f(x)$ and $k = h\, f'(x) + o(h)$, then $k = O(h) = o(1)$ at $h = 0$, so $o(k) = o(h)$. Consequently,

$$\begin{aligned} (g \circ f)(x+h) &= g(y+k) = g(y) + k\, g'(y) + o(k) \\ &= g\big(f(x)\big) + \big(h\, f'(x) + o(h)\big)\, g'(y) + o(h) \\ &= g\big(f(x)\big) + h\, g'\big(f(x)\big) \cdot f'(x) + o(h), \end{aligned}$$

which completes the proof. □

The chain rule is one of the most powerful computational tools in differential calculus. Consider attempting to differentiate $p(x) = (4 + x - x^3)^{11}$. Without the chain rule, the only way to proceed is to multiply out, getting a polynomial of degree 33, then to differentiate using Proposition 8.4. Assuming no mistakes are made, the answer comes out in unfactored form, and factoring it is no mean feat. By contrast, the chain rule gives the factored answer in a single step. Define f and g by $f(x) = 4 + x - x^3$ and $g(y) = y^{11}$. (The use of y is purely for psychological convenience, so we can set $y = f(x)$ in a moment.) The formulas above for the derivative of a polynomial function imply $f'(x) = 1 - 3x^2$ and $g'(y) = 11y^{10}$. Since $p = g \circ f$, the chain rule gives

$$p'(x) = g'\big(f(x)\big) \cdot f'(x) = 11\big(4 + x - x^3\big)^{10}(1 - 3x^2).$$

The chain rule looks particularly compelling in Leibniz notation. If we write $y = f(x)$ and $z = g(y)$, then $z = (g \circ f)(x)$, so

$$\frac{dz}{dx}(x) = \frac{dz}{dy}(y) \cdot \frac{dy}{dx}(x), \quad \text{or even} \quad \frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}.$$

The chain rule may therefore be regarded as a theorem that justifies a certain formal manipulation for quotients of infinitesimals. Lest this interpretation make the result seem obvious ("just cancel the *dy*s"), remember that

- The chain rule looks like cancellation of fractions because we have *denoted* derivatives like fractions, not because they *are* fractions. An infinitesimal like dx is, for us, meaningless in isolation. Logically, it is no more legitimate to "cancel the dy's" than it is to cancel the n's and "deduce" that $\frac{\sin x}{\tan x} = \frac{\text{six}}{\text{tax}}$. In addition, we modified notation (by omitting arguments of functions) in order to make the conclusion look like fraction cancellation.

- The "z" on the left side represents the value of the function $g \circ f$ at x, or the function $g \circ f$ itself. The "z" on the right-hand side of the chain rule represents the value of g at y, or the function g itself. *These z's are usually not the same function!*

Generally, needless confusion results from writing functions in Leibniz notation (as scientists are fond of doing) and using Newtonian derivative notation (as mathematicians are fond of doing); see Exercise 8.4 for a simple example of this pitfall. However, the "cancellation of fractions" interpretation of the chain rule can be a useful mnemonic, provided you remember the fine points just mentioned.

8.2 Derivatives and Local Behavior

If f is differentiable at x for every x in its domain, then f is said to be *differentiable*. In this case, there is a function f', with domain equal to the domain of f and defined for each x by

$$f'(x) = \lim(\Delta_x f, 0) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}. \tag{8.5}$$

The Sign of the Derivative

If f is differentiable at x, then writing $f(x) = a_0$ and $f'(x) = a_1$ for simplicity we have

$$f(x+h) = a_0 + a_1 h + o(h).$$

This condition asserts that f is approximated by a linear function near x; the difference between $f(x+h)$ and a linear function is vanishingly small compared to h.

Suppose that $f'(x) = a_1 > 0$. The dominant non-constant term above is $a_1 h$, which implies the values of f are larger than $f(x)$ in some interval to the right of x, and are smaller than $f(x)$ in some interval to the left of x. Formally, there exists a $\delta > 0$ such that

$$0 < h < \delta \implies f(x-h) < f(x) < f(x+h).$$

This condition is expressed by saying *f is increasing at x*. An analogous argument shows that if $f'(x) < 0$, then (in an obvious sense) f is decreasing at x.

Remark 8.11 If f is increasing at x, it does *not* follow that there exists a $\delta > 0$ such that f is increasing on the interval $(x-\delta, x+\delta)$. The signum function

$$\operatorname{sgn}(x) = \begin{cases} x/|x| & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

is increasing at 0 (read the definition carefully!) but not increasing on any neighborhood of 0. Exercise 8.12 describes an example that is continuous at 0 but is not even non-decreasing on any neighborhood of 0. In Exercise 8.15, you will find a *differentiable* function g with $g'(0) > 0$ that fails to be increasing in any open interval about 0! □

The observations about the sign of the derivative allow us to state and prove an important property related to optimization. By Theorem 5.8, a continuous function $f : [a,b] \to \mathbf{R}$ has a minimum and a maximum. The arguments above show that a point x at which $f'(x) \neq 0$ cannot be an extremum of f.

Theorem 8.12. *Let $f : [a,b] \to \mathbf{R}$ be a continuous function, and let $x_0 \in [a,b]$ be a point at which the minimum or maximum value of f is achieved. Then x_0 is one of the following:*

- *An endpoint of $[a,b]$;*
- *An interior point such that $f'(x_0) = 0$;*
- *An interior point at which $f'(x_0)$ does not exist,*

A point $x \in (a,b)$ where $f'(x) = 0$ is a *critical point* of f. One reason critical points are important is that they are potential locations of the extrema of f. Example 8.13 illustrates the use of Theorem 8.12 in finding extrema.

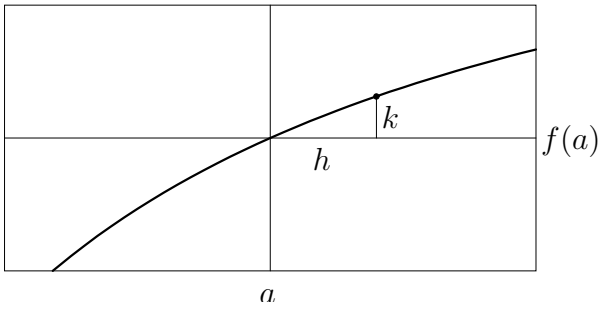


Figure 8.4: Zooming in on a graph.

Tangent Lines and Derivatives

Geometrically, differentiating a function f at a amounts to "zooming in on the graph with factor infinity." To stretch this Leibniz-style metaphor further, the graph of a differentiable function is made up of infinitely many infinitesimal line segments, and the slope of the segment at a is $f'(a)$, the "rise over the run at a." The purpose of this section is to give weight to these remarks.

Let f be a function that is differentiable at a. A line passing through the point $\big(a, f(a)\big)$ is *tangent* to the graph if the slope of the line is $f'(a)$. Intuitively, the tangent line to the graph is an infinitesimal "model" of the graph.

To see why "zooming in with factor infinity at $\big(a, f(a)\big)$" amounts to finding the tangent line at a, consider what zooming in at $\big(a, f(a)\big)$ does to the plane. If h and k denote the horizontal and vertical displacements from the center of magnification (Figure 8.4), then zooming in by a factor of λ maps (h, k) to $(\lambda h, \lambda k)$. This replaces the graph $k = f(a + h) - f(a)$ by the graph $k/\lambda = f(a + h/\lambda) - f(a)$, or

$$k = \left(\frac{f(a + h/\lambda) - f(a)}{(h/\lambda)} \right) \cdot h.$$

If f is differentiable at a, then as $\lambda \to \infty$ the equation above approaches $k = f'(a) \cdot h$, the equation of the tangent line.

In o notation, there is a simpler (but less rigorous) explanation: Since $f(a + h) = f(a) + h\, f'(a) + o(h)$, zooming in with factor infinity kills the negligible term $o(h)$, leaving the equation of the tangent line.

Optimization

If $f : [a, b] \to \mathbf{R}$ is a continuous function, then f achieves a maximum and minimum value, by the extreme value theorem. In practice, it is often desired to *locate* the extreme points of a function, not merely prove their existence. Theorem 8.12, which asserts that extreme points must be endpoints, critical points, or points of non-differentiability, is a useful tool in this situation.

Example 8.13 Suppose we wish to find the rectangle of largest area that has its bottom side on the x axis and is inscribed in the parabola $y = 1 - x^2$, Figure 8.5.

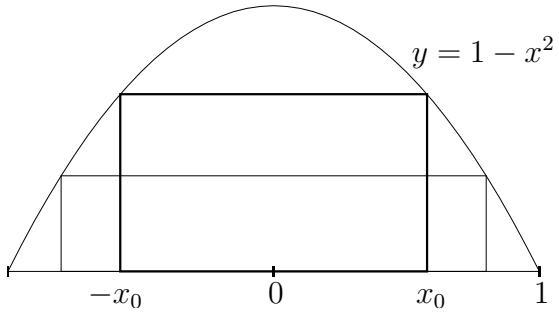


Figure 8.5: The maximum-area rectangle inscribed in a parabola.

If we let $x \geq 0$ be the coordinate of the right side of the rectangle, then the area is $A(x) = 2x(1-x^2) = 2x-2x^3$ for $x \in [0, 1]$. The function A is continuous on a closed, bounded interval, so there is a maximum point by the extreme value theorem. Further, the function A is differentiable at each point of $(0, 1)$, and $A'(x) = 2 - 6x^3 = 2(1 - 3x^2)$. There is only one critical point, $x_0 = 1/\sqrt{3}$, so the extreme points of A must be found in the list 0, x_0, 1. Since $A(0) = A(1) = 0$, while $A(x_0) > 0$, x_0 must be the maximum point! We obtain the maximum area, $A(x_0) = 4/(3\sqrt{3})$, as a fringe benefit. Finding the largest-area rectangle with algebra and geometry alone is not an easy task. □

The argument just given is a process of elimination. First, we know that a maximum point of A exists in $[0, 1]$. Second, we know that if $0 < x < 1$ and $A'(x) \neq 0$, then x is *not* an extreme point. This fact eliminates all but three possibilities, listed above. The endpoints cannot be maximum points, because the area function vanishes at the endpoints and is positive elsewhere. The only remaining possibility is that x_0 is the maximum point.

In other situations, there may be multiple critical points, but Theorem 8.12 is still helpful as long as the function to be optimized is differentiable on (a, b) and continuous on $[a, b]$; as before, the extrema must either be endpoints or critical points, and if there are only finitely many critical points, then the search for extrema is reduced to a finite search.

Example 8.14 Suppose that we want to know the minimum and maximum values of the polynomial $f(x) = x - x^3/6$, subject to $-2 \leq x \leq 3$. First note that f is differentiable on $(-2, 3)$, and that $f'(x) = 1 - x^2/2$, so the critical points of f are $-\sqrt{2}$ and $\sqrt{2}$. Theorem 8.12 guarantees that the extrema must occur in the list -2, $-\sqrt{2}$, $\sqrt{2}$, and 3. Direct calculation gives

$$f(-2) = \frac{2}{3}, \quad f(-\sqrt{2}) = -\frac{2\sqrt{2}}{3}, \quad f(\sqrt{2}) = \frac{2\sqrt{2}}{3}, \quad f(3) = -\frac{3}{2},$$

so the question is reduced to finding the largest and smallest numbers in this list.

Now, $2\sqrt{2} = \sqrt{8} < \sqrt{9} = 3$, so the smallest value is $f(3) = -3/2$; the unique minimum point of f is 3, and the minimum value of f is $-3/2$. Similarly, $1 < \sqrt{2}$, so the largest value is $f(\sqrt{2})$: The unique maximum point is $\sqrt{2}$, and the maximum value of f is $2\sqrt{2}/3$. □

8.3 Continuity of the Derivative

If $f : (a, b) \to \mathbf{R}$ is differentiable, then there is a function $f' : (a, b) \to \mathbf{R}$; however, the function f' is not continuous in general. If f' *is* a continuous function, then f is said to be *continuously differentiable*, or $\mathscr{C}^1$. The set of such functions is denoted $\mathscr{C}^1(a, b)$. For instance, a rational function is $\mathscr{C}^1$ on its natural domain, since the derivative is another rational function with the same domain.

There is a good chance you have never seen a differentiable function with discontinuous derivative. The natural first guess, the absolute value function, is not an example, as it fails to be differentiable at 0 (where the discontinuity "ought to be"). In fact, we must be substantially more devious:

Example 8.15 Let $\psi : \mathbf{R} \to \mathbf{R}$ be a non-constant, differentiable, periodic function. (We construct such functions in Example 9.11 and

Chapter 13.) Define

$$f(x) = \begin{cases} x^2\psi(1/x) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

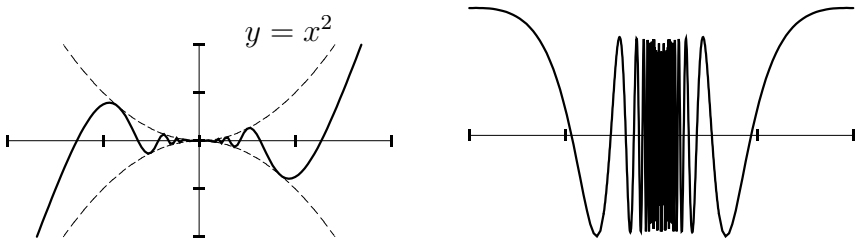


Figure 8.6: A non-$\mathscr{C}^1$ function and its derivative.

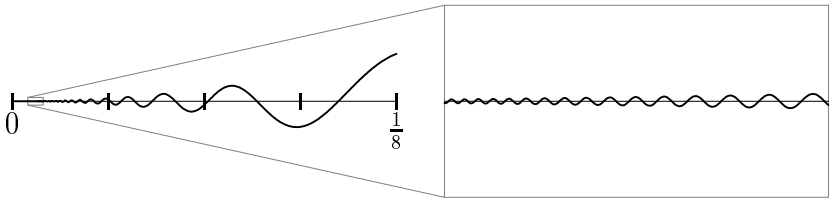
Away from 0, f is obtained by composing and multiplying differentiable functions, and is therefore differentiable by Theorems 8.7 and 8.10:

$$f'(x) = 2x\psi(1/x) - \psi'(1/x) \qquad \text{for } x \neq 0.$$

At $x = 0$ these theorems do not apply (their hypotheses are not satisfied), but the derivative at 0 can be computed from the definition; indeed, $f(h) = O(h^2)$ near $h = 0$, so by the remarks at the end of Example 8.3, $f'(0)$ exists and is equal to 0. In summary, f is differentiable on all of $\mathbf{R}$. It is left to you to verify that f' is not continuous at $x = 0$, see Exercise 8.15. □

It is important to understand exactly what is happening near 0 in Example 8.15. Figure 8.6 is a starting point, but even better it to use a graphing program that can zoom and display at high resolution. Figure 8.6 and the pictures below were drawn using $\psi = \sin$.

If you zoom in at $x = 0$, the graph quickly flattens out into a horizontal line; this reflects that fact that $f'(0) = 0$. However, if you zoom in at a point *close* to 0, the graph first magnifies into an approximation of the graph of ψ before settling down to the tangent line. This reflects the property that the slopes of the tangent lines oscillate infinitely often as $x \searrow 0$.



While f is small in absolute value near 0, its derivative is not. Example 8.15 shows exactly why this can happen: A graph that lies near the horizontal axis can have small oscillations of large slope. In terms of linear mappings, D can take a pair of functions whose difference is small and map them to functions whose difference is large. This should remind you of what a discontinuous function does.

8.4 Higher Derivatives

If $f : (a, b) \to \mathbf{R}$ is differentiable, then the derivative f' is a function with domain $X := (a, b)$, and it makes sense to ask whether or not f' is itself differentiable. If so, we say f is *twice differentiable*; the derivative of f' is denoted $f'' = D^2 f$, and is called the *second derivative* of f. In anticipation of things to come, we also write $f^{(2)}$ for the second derivative of f. The set of twice-differentiable functions is a vector subspace $\mathscr{D}^2(X) \subset \mathscr{F}(X, \mathbf{R})$.

Considerations of continuity apply to second derivatives; a function having *continuous* second derivative is said to be $\mathscr{C}^2$, and the set of all such functions is a subspace. A moment's thought will convince you of the inclusions

$$\mathscr{C}^2(X) \subset \mathscr{D}^2(X) \subset \mathscr{C}^1(X) \subset \mathscr{D}^1(X) \subset \mathscr{C}(X) \subset \mathscr{F}(X, \mathbf{R}).$$

As you might guess, the pattern continues off to the left; the vector subspace of k *times continuously differentiable functions* is defined by

$$\mathscr{C}^k(X) = \{f \in \mathscr{F}(X, \mathbf{R}) \mid f^{(k)} \text{ exists and is continuous}\}.$$

In this book we are not so interested in these spaces, though we will meet several members of their intersection, the space of *smooth* functions:

$$\mathscr{C}^\infty(X) = \bigcap_{k=1}^{\infty} \mathscr{C}^k(X) = \{f \in \mathscr{F}(X, \mathbf{R}) \mid f^{(k)} \text{ exists for all } k \text{ in } \mathbf{N}\}.$$

Polynomials and rational functions are smooth (on their natural domain), as is the natural logarithm, whose derivative is rational. The function f of Example 8.15 is smooth except at 0, but of course is not even $\mathscr{C}^1$ on $\mathbf{R}$. Exercises 8.9 and 8.11 give more examples of non-smooth functions.

Higher Order Differences

There is not much we can say about higher order derivatives at this stage, because we have not rigorously established certain "obvious" properties of the derivative (e.g., if $f' > 0$, then f is increasing). As with "obvious" properties of continuous functions, familiar properties of differentiable functions are more subtle than they first appear, and are not actually true unless some care is taken with hypotheses! The technical tool needed to study derivatives is the "mean value theorem", the subject of Chapter 9.

However, elementary algebra gives us some idea of the information encoded in the first and higher derivatives of a function. For the rest of the section, let f be a real-valued function whose domain is an interval I; all points are assumed to be elements of I without further mention.

The difference quotient

$$\frac{\Delta y}{\Delta x}(a, b-a) = \frac{f(b) - f(a)}{b - a}$$

measures the rate of change of f on the interval $[a, b]$. In real experiments, difference quotients are all one ever knows, because it is not possible (or even philosophically meaningful) to collect data for all points in the domain. Instead, scientists assume there exists a mathematical model (unknown at the outset), and that measured data arises as outputs of the model (up to experimental error).

At least two measurements are required to determine whether a function is (on the average) increasing or decreasing. Two measurements of f correspond to a single measurement of f', which is computed by sampling f at two infinitesimally separated points.

Now suppose we want to measure *how fast the rate of change is varying*. The rate of change is f', which varies at rate f''. We need two measurements of f', or *three* measurements of f. Imagine waiting to cross a busy street, looking left and right (like someone at a tennis match) to see if cars are coming. You must make two observations to

determine how fast a vehicle is traveling. In addition, if you see a car approaching from the left, then observe that no one is coming from the right, it is still prudent to look left again, to see whether or not the oncoming car is accelerating. If after the third glance the car has traveled much further than it did between your first two sightings, you should re-evaluate whether it is safe to cross.[4]

As another example, consider a company whose net worth is $V(t)$ dollars at time t measured in months from January, 2000 (say). We must see at least two quarterly reports (i.e., obtain two values of V) before we can determine whether the company is earning or losing money, and must see at least three reports to know whether earnings are up or down. In business circles, a company is often considered to be "losing money" if the net worth of the company is increasing, but the rate at which the net worth is increasing is decreasing, i.e., if the *second derivative* of the net worth is negative.

Exercises

Exercise 8.1 Express Proposition 8.4 in Leibniz notation, and explain how the result is useful as a tool in formal manipulations. ◇

Exercise 8.2 Prove Proposition 8.8. ◇

Exercise 8.3 Let n be a positive integer, and consider the finite geometric series

$$\sum_{k=0}^{n} x^k = 1 + x + x^2 + \cdots + x^n = \frac{x^{n+1} - 1}{x - 1} \qquad \text{if } x \neq 1.$$

(a) Use Theorem 8.7 and Proposition 8.8 to differentiate this equation when $x \neq 1$.

(b) Use part (a) to find a closed-form expression for the series

$$\sum_{k=1}^{n} kx^k = x + 2x^2 + 3x^3 \cdots + nx^n, \qquad x \neq 1.$$

[4]This advice is distilled from an incident in which the author was nearly hit by a speeding cab at the intersection of Bloor and St. George streets in Toronto.

(c) Continue in the same vein, deducing that

$$\sum_{k=1}^{n} k^2 x^k = \frac{n^2 x^{n+2} - (2n^2 + 2n - 1)x^{n+1} + (n+1)^2 x^n - (x+1)}{(x-1)^3}$$

for $x \neq 1$.

The technique of integrating or differentiating a known sum is a powerful trick in the right circumstances. ◇

Exercise 8.4 Suppose $y = x^2$ and $z = y^2$, so that $z = x^4$. Then $z'(y) = 2y$, and when $x = 1$ we get $z'(1) = 2$. However, $z'(x) = 4x^3$, so at $x = 1$ we have $z'(1) = 4$. Therefore $2 = 4$. What is wrong? ◇

Exercise 8.5 Prove that among all rectangles of perimeter $P > 0$, there exists one of largest area, and find its dimensions. Solve this problem both with calculus and by pure algebra (completing the square). ◇

Exercise 8.6 Prove that among all rectangles of area A, there exists one of smallest perimeter, cf. Exercise 8.5. This problem is much easier to do with pure algebra than with calculus, because you cannot use the extreme value theorem to deduce existence of a minimum. The moral is that calculus is not always the best technique for optimization. ◇

Exercise 8.7 Find the dimensions of the rectangle of largest area inscribed in a half-disk of radius r; you may assume that one side of the rectangle lies along the diameter. ◇

Exercise 8.8 Consider the family of rectangles whose lower left corner lies at the origin, whose upper right corner lies on the graph $y = 1/(1+x^2)$, and whose sides are parallel to the coordinate axes. Prove that there exists a rectangle of largest area in this family, and find its dimensions. ◇

Continuity of Derivatives

Exercise 8.9 Let k be a positive integer, and define $f : \mathbf{R} \to \mathbf{R}$ by $f(x) = x^k|x|$. Find the derivative of f, and prove that f is $\mathscr{C}^k$ but is not $(k+1)$ times differentiable. In other words, the inclusion $\mathscr{C}^k(\mathbf{R}) \subset \mathscr{D}^{k+1}(\mathbf{R})$ is proper.
Suggestion: Do induction on k. ◇

Exercise 8.10 Let $f : \mathbf{R} \to \mathbf{R}$ be differentiable but not $\mathscr{C}^1$, and put

$$F(x) = \int_0^x f(t)\,dt.$$

Prove that F is twice-differentiable, but is not $\mathscr{C}^2$. ◇

Exercise 8.11 Let $k > 1$ be an integer. Continuing the previous exercise, prove that there exists a function f that is k times differentiable, but not $\mathscr{C}^k$. In other words, the inclusion $\mathscr{D}^k(\mathbf{R}) \subset \mathscr{C}^k(\mathbf{R})$ is proper. ◇

Exercise 8.12 Define $f : \mathbf{R} \to \mathbf{R}$ by

$$f(x) = \begin{cases} x \text{ if } x \in \mathbf{Q} \\ 2x \text{ if } x \notin \mathbf{Q} \end{cases}$$

Show that f is increasing at 0, but that there does not exist $\eta > 0$ such that f is increasing on the open interval $(-\eta, \eta)$. ◇

Exercise 8.13 Let $\psi : \mathbf{R} \to \mathbf{R}$ be a non-constant, differentiable, periodic function whose derivative varies between -1 and 1.

(a) For n a positive integer, let $f_n : \mathbf{R} \to \mathbf{R}$ be defined by $f_n(x) = \frac{1}{n}\psi(n^2 x)$. Find

$$\max_{x \in \mathbf{R}} |f_n(x)| \quad \text{and} \quad \max_{x \in \mathbf{R}} |f_n'(x)|.$$

(b) Given an example of a differentiable function $f : \mathbf{R} \to \mathbf{R}$ such that $\lim(f, +\infty)$ exists but $\lim(f', +\infty)$ does not exist.

Part (a) is meant to suggest an idea for constructing f, though of course your answer for (b) should not depend on n. ◇

Exercise 8.14 Let $\psi : \mathbf{R} \to \mathbf{R}$ be as in Exercise 8.13, and define

$$f(x) = \begin{cases} x^2\psi(1/x) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

We found the derivative of f in Example 8.15. Prove that f' is discontinuous at 0. It may help to review Chapter 4. ◇

Exercise 8.15 Let f be as in Exercise 8.14, and let $g(x) = f(x)+(x/2)$. Prove that $g'(0) > 0$, but there does not exist an open interval about 0 on which g is increasing. Why doesn't this contradict the fact that "a function with positive derivative is increasing"? ◇

Chapter 9

The Mean Value Theorem

The results of the last chapter depend mostly on a function being differentiable at a single point, and are therefore of a pointwise, or at most local, nature. In this chapter, we link together the machinery of differentiability with the global theorems on continuity from Chapter 5. The mean value theorem equates, under suitable hypotheses, the average rate of change of f over an interval $[a, b]$ and the instantaneous rate of change of f at some point of (a, b). The result allows us to pass between a collection of pointwise information and global information, and is rightly regarded as a technical foundation stone of the calculus.

9.1 The Mean Value Theorem

In Chapter 5, we assumed that f was continuous on $[a, b]$ and deduced global properties of f. Here we assume, in addition, that $f : [a, b] \to \mathbf{R}$ is differentiable on the open interval (a, b). (Equivalently, if redundancy upsets you, f is differentiable on (a, b), and continuous at a and b.) A typical example is the function $f(x) = \sqrt{1 - x^2}$ on $[-1, 1]$, whose graph is the upper half of the unit circle in the plane.

Theorem 9.1. *Let $f : [a, b] \to \mathbf{R}$ be a continuous function that is differentiable on (a, b). Then there exists an $x_0 \in (a, b)$ such that*

$$f'(x_0) = \frac{f(b) - f(a)}{b - a}.$$

In words, there is a point in (a, b) at which the instantaneous rate of change of f is equal to the average rate of change of f over the interval $[a, b]$. Figure 9.1 depicts the conclusion in a simple (but representative)

situation. The conclusion is quite plausible when phrased in terms of speed and distance: If on a car trip you cover 60 miles in a certain one-hour period of time, then at some instant during that hour your speed must have been exactly 60 miles per hour. Of course, this proves nothing, because real distances and speeds do not correspond exactly with real numbers and functions, but it's a good way of remembering the theorem's conclusion.

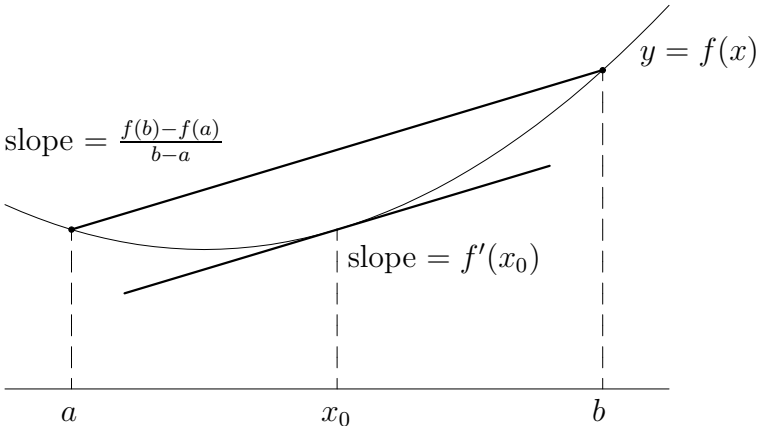


Figure 9.1: The conclusion of the mean value theorem.

Proof. The proof of the mean value theorem breaks conceptually into two steps. The first, called *Rolle's theorem*, treats the special case where the function values at the endpoints are equal, and uses the extreme value theorem and Theorem 8.12. The second step reduces the theorem to Rolle's theorem by an algebraic trick.

Assume first that $f(a) = f(b)$, so the average rate of change is 0. We wish to show that f has a critical point. By the extreme value theorem, there exist points $x_{\min}$ and $x_{\max} \in [a, b]$ such that

$$f(x_{\min}) \leq f(x) \leq f(x_{\max}) \quad \text{for all } x \in [a, b].$$

(These points are not in general unique.) Suppose first that at least one of $x_{\min}$ and $x_{\max}$ is in (a, b), and call it x_0. By Theorem 8.12, $f'(x_0) = 0$ and we are done. The only other possibility is that each of the points $x_{\min}$ and $x_{\max}$ is an endpoint of $[a, b]$. But since $f(a) = f(b)$, this means f is a constant function, and then $f'(x_0) = 0$ for *every* point $x_0 \in (a, b)$. This proves Rolle's theorem.

Next consider the function $g : [a, b] \to \mathbf{R}$ defined by "linearly adjusting" the endpoint values of f so they are equal:

$$g(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a).$$

The function g is continuous on $[a, b]$ and differentiable on (a, b), as a sum of functions that have these properties. Direct calculation shows that $g(a) = f(a) = g(b)$, so g satisfies the hypotheses of Rolle's theorem. Consequently there exists an $x_0 \in (a, b)$ such that $g'(x_0) = 0$. But again, direct calculation gives

$$g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a} \quad \text{for all } x \in (a, b),$$

and the theorem follows. □

We are now in a position to derive some easy but important consequences. You should bear in mind how difficult the following theorems are to prove without the mean value theorem.

9.2 The Identity Theorem

A constant function has derivative identically zero. Conversely, it seems reasonable that a function whose derivative vanishes everywhere must be a constant function. If the domain is an interval of real numbers, this is indeed true. However, you should note well that the proof is impossible without the mean value theorem.

Theorem 9.2. *Let f and g be differentiable functions on an interval I. If $f' = g'$, then there exists a real number c such that $f(x) = g(x) + c$ for all $x \in I$.*

It is, by the way, crucial that the domain be an interval of real numbers. Consider the function sgn $x = x/|x|$, defined for all $x \neq 0$; the derivative vanishes identically on the domain, but sgn is not constant.

Proof. By consideration of the differentiable function $h = f - g$, it suffices to show that if $h'(x) = 0$ for all $x \in I$, then h is a constant function. We prove the contrapositive: If h is non-constant on an interval, then h' is not identically zero.

Suppose h is a non-constant function on I, and pick a, $b \in I$ so that $h(a) \neq h(b)$. By the mean value theorem, there is an x_0 between a and b such that

$$h'(x_0) = \frac{h(b) - h(a)}{b - a}.$$

The right-hand side is non-zero, so h' is not identically zero. □

The identity theorem is most often used to show that two functions are equal. If f and g are differentiable functions that have the same derivative on some interval, then f and g differ by a constant on that interval. If, in addition, $f(x_0) = g(x_0)$ for some x_0, then f and g are equal in the interval. We will put this technique to good use, as we have many interesting ways of procuring pairs of functions that have the same derivative.

Monotonicity and the Sign of the Derivative

The identity theorem tells us what to expect when a function has vanishing derivative. Interesting conclusions can be drawn when the derivative is everywhere positive, or everywhere negative. You should compare the result below with the observations of Section 8.2, and with Exercise 8.15.

Theorem 9.3. *If $f : (a, b) \to \mathbf{R}$ is differentiable, and if $f'(x) > 0$ for all $x \in (a, b)$, then f is increasing on (a, b).*

Proof. Under the assumptions of the theorem, if $a < x < y < b$ then there is an $x_0 \in (x, y)$ such that

$$\frac{f(y) - f(x)}{y - x} = f'(x_0).$$

By hypothesis, $f'(x_0) > 0$, and since $x < y$, it follows that $f(x) < f(y)$. An entirely analogous argument shows that if $f'(x) < 0$ for all x in some interval, then f is decreasing on the interval. □

The Exponential Function

To demonstrate the power of the theorems just proven, here is a short digression that shows how properties of a function can be studied without having a concrete representation of the function.

The equation $f' = f$ is an example of an *ordinary differential equation*, or ODE for short. The unknown f is a differentiable function whose domain is some unspecified open interval. It is not obvious whether this differential equation has any "interesting" solutions (the zero function is an "uninteresting" solution), and if so, how many solutions it has. We can, nonetheless, determine consequences of the differential equation that tell us properties of any solutions that may exist.

In Example 9.14, we will prove that there exists a non-vanishing differentiable function $\exp : \mathbf{R} \to \mathbf{R}$, the *natural exponential function*, such that $\exp' = \exp$ and $\exp(0) = 1$. For the rest of this section, we assume the existence of exp. Logically, there is no problem, since we are not deducing the existence of exp, but are instead deducing properties that exp must possess. Our knowledge about exp originates with the fact that it solves the differential equation $f' = f$, as we shall see now.

Proposition 9.4. *Let $f : \mathbf{R} \to \mathbf{R}$ satisfy $f' = f$. Then $f(x) = f(0)\exp(x)$ for all $x \in \mathbf{R}$. In particular, if $f' = f$ and $f(0) = 1$, then $f = \exp$.*

Proof. Because exp is differentiable and nowhere-vanishing, the function $q = f/\exp$ is differentiable. The quotient rule implies

$$q' = \frac{\exp\, f' - f\, \exp'}{\exp^2} = \frac{f' - f}{\exp},$$

which vanishes identically because $f' = f$. By the identity theorem, q is a constant function on $\mathbf{R}$, and evaluating at 0 shows that

$$\frac{f(x)}{\exp(x)} = q(x) = q(0) = \frac{f(0)}{\exp(0)} = f(0)$$

for all x, so $f = f(0)\exp$ as claimed. □

Proposition 9.5. *Let k be a real number. If $f : \mathbf{R} \to \mathbf{R}$ is a differentiable function satisfying $f' = kf$ and $f(0) = 1$, then $f(x + y) = f(x)f(y)$ for all x, $y \in \mathbf{R}$*

Proof. The function $g(x) = \exp(kx)$ solves the ODE $g' = kg$ by the chain rule, and satisfies the initial condition $g(0) = 1$. The proof of Proposition 9.4 implies that this is the *only* such function. Consequently, it is enough to show that

$$\exp(x + y) = \exp(x)\exp(y) \qquad \text{for all } x,\, y \in \mathbf{R}.$$

Fix y, and consider the function $f : \mathbf{R} \to \mathbf{R}$ defined by $f(x) = \exp(x + y)$. By the chain rule, f is differentiable, and $f' = f$. Since $f(0) = \exp(y)$, Proposition 9.4 implies $\exp(x+y) = \exp(x)\exp(y)$ for all x. □

The special case $\exp(x)\exp(-x) = \exp(x - x) = \exp(0) = 1$ says that exp is nowhere-vanishing. By continuity, $\exp(x) > 0$ for all real x. From the defining property $\exp' = \exp$, we deduce that exp is an increasing function. The number $e := \exp(1) > \exp(0) = 1$ is a fundamental constant of mathematics. Though we can say little at present about the numerical value of e, we *can* justify the name "exponential function."

Corollary 9.6. $\exp(r) = e^r$ *for all rational* r.

Proof. By induction, $\exp(p) = e^p$ for all $p \in \mathbf{N}$. As noted above, Proposition 9.5 implies $\exp(p)\exp(-p) = 1$ for all p, so $\exp(-p) = 1/\exp(p) = e^{-p}$. Finally, if $q \in \mathbf{N}$, then

$$\begin{aligned}\big[\exp(p/q)\big]^q &= \exp(p/q) \cdot \overset{q}{\cdots} \cdot \exp(p/q) \\ &= \exp(p/q + \overset{q}{\cdots} + p/q) = \exp(p) = e^p,\end{aligned}$$

so $\exp(p/q) = \sqrt[q]{e^p} = e^{p/q}$. □

On the basis of the corollary, it is reasonable to *define* $e^x = \exp(x)$ for all $x \in \mathbf{R}$. Proposition 9.5 is the familiar law

$$e^{x+y} = e^x e^y \qquad \text{for all } x,\, y \in \mathbf{R}.$$

Remember, we have not yet shown that exp exists, but we have deduced a number of properties it must have, assuming only that $\exp' = \exp$ and $\exp(0) = 1$.

The Intermediate Value Property of the Derivative

Let $f : I \to \mathbf{R}$ be a differentiable function on an interval I. A theorem of Darboux[1] asserts that f' has the intermediate value property; in particular, a discontinuity of a derivative must be wild. The function of Example 8.15 was not especially pathological!

Theorem 9.7. *Let f be differentiable on some open interval containing $[a, b]$. If c is a real number between $f'(a)$ and $f'(b)$, then there exists $x_0 \in (a, b)$ such that $f'(x_0) = c$.*

[1] dar BOO

Proof. Assume without loss of generality that $f'(a) < f'(b)$, and consider the differentiable function g defined by $g(x) = f(x) - cx$. Since g is continuous on $[a, b]$, g achieves its minimum at some point $x_0 \in [a, b]$. However, $g'(a) = f'(a) - c < 0$, so g is decreasing at a. This means the minimum of g is not achieved at a. Similarly, $g'(b) = f'(b) - c > 0$, so g is increasing at b, which means b is not a minimum of g. The minimum value of g must therefore be attained at some point $x_0 \in (a, b)$, and by Theorem 8.12 $g'(x_0) = 0$, or $f'(x_0) = c$. □

Corollary 9.8. *Let $f : I \to \mathbf{R}$ be a differentiable function on an interval containing $[a, b]$. If f' is non-vanishing in the open interval (a, b), then f is strictly monotone—hence invertible—in the closed interval $[a, b]$.*

Proof. Suppose, without loss of generality, that $f'(x) > 0$ for some $x \in (a, b)$. Darboux' theorem implies f' is positive everywhere in the interval, for if f' were negative somewhere it would have to vanish somewhere. Theorem 9.3 implies f is strictly increasing on the open interval (a, b). Finally, if $x \in (a, b)$, then

$$\frac{f(x) - f(a)}{x - a} > 0$$

just as in the proof of Theorem 9.2, which implies $f(x) > f(a)$. Similarly, $f(x) < f(b)$. □

Corollary 9.8 gives a sufficient criterion for invertibility of a function. For rational functions, this condition is often extremely easy to check, and indeed is usually the simplest means of proving a function is invertible on some interval.

Example 9.9 If $f(x) = x - x^3/3$ for $x \in \mathbf{R}$, then $f'(x) = 1 - x^2$, and the critical points are -1 and 1. The derivative—a polynomial function—is continuous, so Corollary 9.8 implies f is one-to-one on each of the intervals $(-\infty, -1]$, $[-1, 1]$, and $[1, \infty)$. In fact, $f'(x) > 0$ iff $|x| < 1$, so f is decreasing on each of the unbounded intervals, and is increasing on $[-1, 1]$.

These intervals share endpoints, but there is no contradiction, as should be clear from Figure 9.2. □

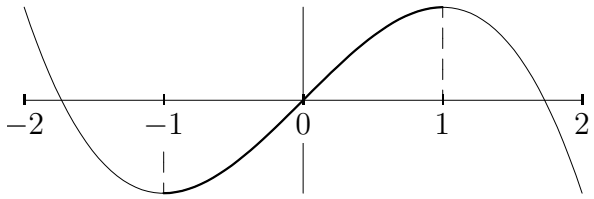


Figure 9.2: Intervals on which a polynomial is monotone.

Patching

In applications it can be desirable to present a function by giving two or more formulas that hold on abutting intervals, e.g.,

$$\psi(x) = 4x(1-|x|) = \begin{cases} 4x(1+x) & \text{if } -1 \le x \le 0 \\ 4x(1-x) & \text{if } 0 < x \le 1 \end{cases} \tag{9.1}$$

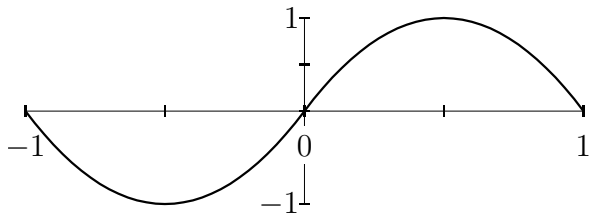


Figure 9.3: The function ψ.

We would like to know when such a "patched" function is differentiable at the point(s) where the formulas "join". The next theorem gives a sufficient criterion that is adequate for many applications.

Theorem 9.10. *Let f be a function that is continuous at x_0 and differentiable in some deleted interval about x_0. If $\lim(f', x_0)$ exists and is equal to ℓ, then f is differentiable at x_0, and $f'(x_0) = \ell$.*

In particular, f' is a continuous function at x_0. As reasonable as this result may seem, the proof requires the mean value theorem. It is instructive to attempt a "naive" proof; the snag is that $\lim(f', x_0)$ involves a double limit, which the theorem interchanges:

$$\lim_{x\to x_0}\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} \stackrel{?}{=} \lim_{h\to 0}\lim_{x\to x_0}\frac{f(x+h)-f(x)}{h}$$

Proof. By hypothesis, there exists a $\delta > 0$ such that on the closed interval $[x_0 - \delta, x_0 + \delta]$ the function f is continuous, and differentiable except possibly at x_0. In particular, f satisfies the hypotheses of the mean value theorem on each of the intervals $[x_0 - \delta, x_0]$ and $[x_0, x_0 + \delta]$. For each number h with $0 < |h| < \delta$, there exists an $x_h \in (x_0, x_0 + h)$ such that

$$f'(x_h) = \frac{f(x_0 + h) - f(x_0)}{h}.$$

By construction, $x_h \to x_0$ as $h \to 0$; taking limits gives $\lim(f', x_0) = f'(x_0)$, as claimed. □

Writing a formal ε-δ proof of the last step is a good exercise.

Example 9.11 Let $\psi : \mathbf{R} \to \mathbf{R}$ be the 2-periodic function whose restriction to $[-1, 1]$ is given by (9.1). For $x \in (0, 1)$ we have $\psi'(x) = 4 - 8x$, while on $(-1, 0)$ we have $\psi'(x) = 4 + 8x$. Using Theorem 9.10, we find that ψ is differentiable at 0, and $\psi'(0) = 4$. Similarly, also using periodicity, we find that ψ is differentiable at ± 1, and $\psi'(\pm 1) = -4$. Since ψ is differentiable over an entire period, it is differentiable on $\mathbf{R}$. In particular, we have constructed a non-constant, periodic function of class $\mathcal{C}^1$.

We call ψ the *pseudo-sine* function. □

9.3 Differentiability of Inverse Functions

Differential calculus provides an effective tool, the sign of the derivative, for determining whether a function is monotone. For differentiable functions whose domain is an interval, monotonicity is *equivalent* to invertibility. We now turn to the question of whether an inverse function is itself differentiable, and if so, how to calculate the derivative.

Let $f : I \to \mathbf{R}$ be a one-to-one function whose domain is an open interval, and let $J = f(I)$ be the image. There is a function $g : J \to I$ such that

$$\begin{aligned} g\big(f(x)\big) &= x \qquad \text{for all } x \in I, \\ f\big(g(y)\big) &= y \qquad \text{for all } y \in J. \end{aligned} \tag{9.2}$$

In other words, for $x \in I$, the equations $y = f(x)$ and $x = g(y)$ are equivalent. Replacing f by $-f$ if necessary, we may as well assume f is increasing.

Theorem 9.12. *Let $f : I \to \mathbf{R}$ be one-to-one and differentiable on the interval I, and let $x_0 \in I$. The function $g = f^{-1}$ is differentiable at $y_0 = f(x_0)$ iff $f'(x_0) \neq 0$, and in this event*

$$g'(y_0) = \frac{1}{f'(x_0)}.$$

Proof. Assume first that $g = f^{-1}$ is differentiable at $y_0 = f(x_0)$. The chain rule applied to the first of (9.2) implies $1 = g'(y_0)f'(x_0)$. We deduce that

$$g'(y_0) = g'\big(f(x_0)\big) = \frac{1}{f'(x_0)} \qquad \text{if } f'(x_0) \neq 0,$$

and that if $f'(x_0) = 0$, then g is not differentiable at y_0. We therefore know what the derivative of f^{-1} must be, *provided the derivative exists.* The Leibniz version of this equation is the natural-looking equation

$$1 = \frac{dy}{dx} \cdot \frac{dx}{dy},$$

with the usual proviso that the xs are not the same, and neither are the ys.

To prove that g really is differentiable when $f' \neq 0$, observe that there exists an $\eta > 0$ such that $(x_0 - \eta, x_0 + \eta) \subset I$, and that for $|h| < \eta$ we may write $f(x_0 + h) - f(x_0) = k$, where k is uniquely determined by h, see Figure 9.4. Rewriting this as $f(x_0 + h) = f(x_0) + k = y_0 + k$ and applying g to both sides,

$$\frac{g(y_0 + k) - g(y_0)}{k} = \frac{h}{f(x_0 + h) - f(x_0)}.$$

If $f'(x_0) > 0$, then the right-hand side has a limit as $h \to 0$, namely $1/f'(x_0)$, while the left-hand side is a Newton quotient for $g'(y_0)$. This proves that if $f'(x_0) \neq 0$, then $g = f^{-1}$ is differentiable at $f(x_0) = y_0$. □

Example 9.13 Let q be a positive integer, and let $f(x) = x^q$ for $x > 0$. The function f is increasing (hence invertible) and differentiable, with derivative $f'(x) = qx^{q-1} > 0$. The inverse function is the qth root function, $g(y) = y^{1/q}$. By Theorem 9.12, g is differentiable at $y = x^q$, and

$$g'(y) = \frac{1}{f'(x)} = \frac{1}{qx^{q-1}} = \frac{1}{q}y^{(1/q)-1}.$$

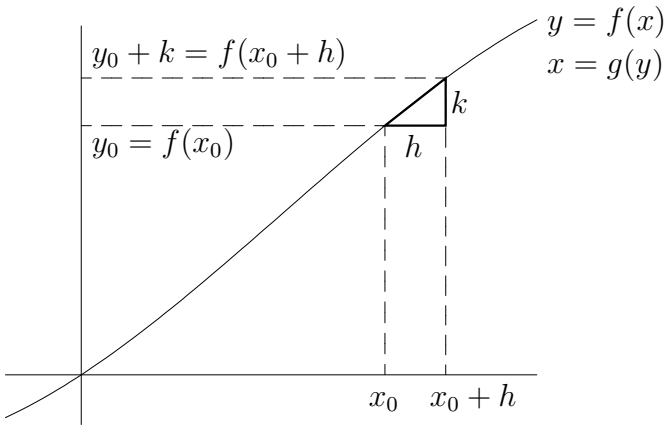


Figure 9.4: The difference quotient of an inverse function.

Exercise 9.4 extends this result to power functions with arbitrary rational exponent. □

Example 9.14 The natural logarithm (Exercise 7.17) is defined by

$$\log x = \int_1^x \frac{1}{t}\,dt, \qquad t > 0.$$

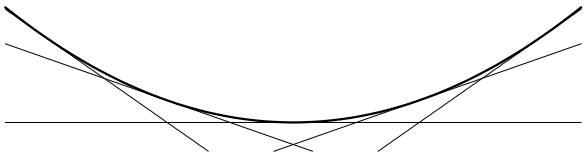
The image of log is $\mathbf{R}$. Theorem 8.9 implies that log is differentiable, and that $\log' x = 1/x$ for $x > 0$. Corollary 9.8 implies log is increasing, hence invertible, a fact we also knew from Exercise 7.17. Let $\exp : \mathbf{R} \to (0, \infty)$ denote the inverse function. Clearly $\log 1 = 0$, so $\exp(0) = 1$, and for each $x > 0$ we have

$$\exp'(x) = \frac{1}{\log'[\exp(x)]} = \frac{1}{1/\exp(x)} = \exp(x).$$

Thus, as we claimed earlier, there exists a differentiable function that is equal to its own derivative on $\mathbf{R}$ and takes the value 1 at 0. □

9.4 The Second Derivative and Convexity

Let $f : [a, b] \to \mathbf{R}$ be continuous, and assume f is twice-differentiable on (a, b). The value $f'(x)$ may be interpreted as the slope of the line tangent to the graph of f at x, and $f''(x)$ is the instantaneous rate of change of the slope as x varies. If $f'' > 0$ on some interval, geometric intuition says the graph of f should be "convex" or "concave up":



The aim of this section is to define "convexity" precisely, and to prove that a $\mathscr{C}^2$ function with positive second derivative is convex is this sense.

A set $R \subset \mathbf{R}^2$ is said to be *convex* if, for all points p_1 and p_2 in R, the segment joining p_1 to p_2 lies entirely within R. To express this criterion algebraically, note that the segment joining $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ is the set of points of the form

$$(1-t)p_1 + tp_2 = \big((1-t)x_1 + tx_2, (1-t)y_1 + ty_2\big) \qquad \text{for } t \in [0,1].$$

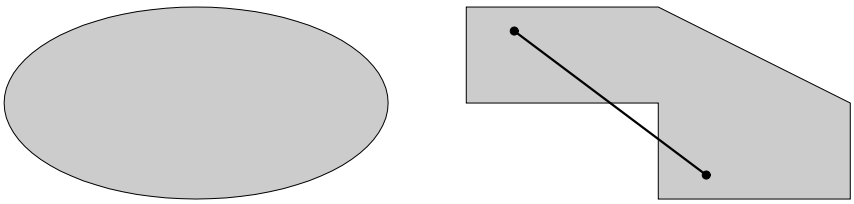


Figure 9.5: Convex and non-convex sets.

Convex Functions

Definition 9.15 Let $I \subseteq \mathbf{R}$ be an interval. A function $f : I \to \mathbf{R}$ is *convex* if the "region above the graph",

$$\Gamma_f^+ := \{(x,y) \mid x \in I,\ f(x) \leq y\},$$

is a convex set in the plane.

A function f is *concave* if $-f$ is convex, i.e., if the "region below the graph",

$$\Gamma_f^- := \{(x,y) \mid x \in I,\ y \leq f(x)\},$$

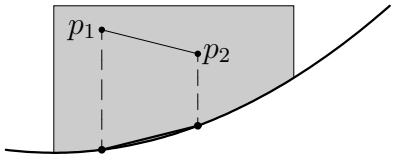
is a convex set in the plane. The terms "concave up" and "concave down" are often used with the same (respective) meanings.

To test the region above the graph of f for convexity, it is enough to choose the points p_i to lie on the graph:

Proposition 9.16. *A function $f : I \to \mathbf{R}$ is convex iff every secant line is contained in Γ_f^+.*

Proof. The "only if" direction is obvious.

For the "if" implication, the idea is that if p_1 and p_2 are points of R, then the segment joining them lies above the secant line obtained by vertical projection:



But the secant line is contained in R by hypothesis, so the segment joining p_1 and p_2 is also contained in R. An algebraic proof is left to you as a translation exercise. □

Just as the sign of the first derivative is related to monotonicity via the mean value theorem, the sign of the second derivative is related to convexity, and the mean value theorem provides the link between infinitesimal information (f'') and finite differences (convexity).

Lemma 9.17. *Suppose f is twice-differentiable on (a, b), $f(x_1) = f(x_2) = 0$ for some $x_1 < x_2$, and that $f'' \geq 0$ on $[x_1, x_2]$. Then $f \leq 0$ on $[x_1, x_2]$.*

Proof. As with our monotonicity results, the contrapositive is more natural to prove: If there exists an x in (x_1, x_2) with $f(x) > 0$, then $f''(z) < 0$ for some z in (x_1, x_2), Figure 9.6.

Applying the mean value theorem to f on the interval $[x_1, x]$, we deduce that there exists $z_1 \in (x_1, x)$ such that

$$f'(z_1) = \frac{f(x) - f(x_1)}{x - x_1} = \frac{f(x)}{x - x_1} > 0.$$

Similarly, there is a point $z_2 \in [x, x_2]$ with $f'(z_2) < 0$.

Now, applying the mean value theorem to f' on $[z_1, z_2]$, we find that

$$f''(z) = \frac{f'(z_2) - f'(z_1)}{z_2 - z_1} < 0$$

for some $z \in (z_1, z_2)$. □

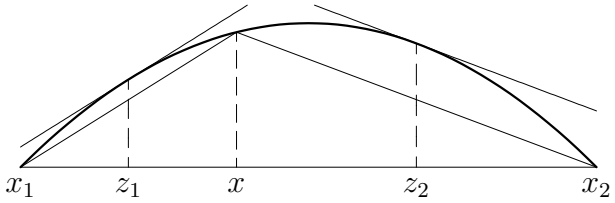


Figure 9.6: Determining the sign of f'' from the value of f.

There is clearly a version of Lemma 9.17 with inequalities reversed. ("If $f'' \leq 0$, then $f \geq 0$.") Further, if f'' is continuous and there is strict inequality in the hypothesis (i.e., $f''(x) < 0$ for some x) then we get strict inequality in the conclusion, since a continuous function that is positive at one point is positive on an interval.

Armed with Lemma 9.17, we can characterize convexity of $\mathscr{C}^2$ functions in terms of the second derivative.

Theorem 9.18. *Suppose f is $\mathscr{C}^2$ on an interval I. For each closed interval $[a,b] \subset I$, f is convex on $[a,b]$ iff $f'' \geq 0$ on (a,b).*

Proof. By Proposition 9.16, it suffices to show that the following conditions are equivalent:

- $f'' \geq 0$ on I.
- "The graph of f lies below every secant line." Formally, for all a, b in I, the segment joining $p_1 = \big(a, f(a)\big)$ and $p_2 = \big(b, f(b)\big)$ lies in Γ_f^+.

Suppose the first condition holds. Fix elements $a < b$ in I, let ℓ be the linear polynomial whose graph is the secant line, and introduce the $\mathscr{C}^2$ function $g = f - \ell$. Direct calculation shows that $g'' = f''$, and that $g(a) = g(b) = 0$. Lemma 9.17 implies $g \leq 0$ on $[a,b]$. But this says $f \leq \ell$ on $[a,b]$, which is what we wanted to prove. Note that we have used twice-differentiability, but not continuity of f''.

Conversely, suppose the first condition above fails: There exists z in (a,b) with $f''(z) < 0$. By continuity of f'', there exists a $\delta > 0$ such that $f'' < 0$ on the interval $[z-\delta, z+\delta]$. As before, let ℓ be the secant line, and let $g = f - \ell$. An obvious modification of Lemma 9.17 (mentioned above) says that $g > 0$ at some point of $[z-\delta, z+\delta] \subset (a,b)$. This means the second condition also fails: There is a secant line that is not contained in Γ_f^+. □

The proof gives more specific information that was stated in the theorem. For example, if $f''(x) > 0$ for some x, then all secant lines in some interval about x lie "strictly above" the graph of f. All these statements have obvious modifications for functions whose second derivative is non-positive. Note also that a function can be convex without being twice-differentiable. The absolute value function is convex on $\mathbf{R}$, but is not even once differentiable, while Exercise 9.15 shows that even a discontinuous function can be convex.

Derivatives and Graphing

Throughout this section, we assume that f is a $\mathscr{C}^2$ function whose domain is an interval of real numbers that contains $[a, b]$. The first and second derivatives can be used to obtain geometric information about the graph of a function. Graphing calculators have led to a de-emphasis on manual graphing techniques, but technical knowledge is still useful, especially for pathological functions that are not well-handled by a computer.

By Darboux' theorem and the mean value theorem, if $f' \neq 0$ in (a, b), then f is monotone in $[a, b]$. If f has only finitely many critical points, a very rough graph can be sketched by plotting the points $\big(x, f(x)\big)$ for each critical point x, then "connecting the dots"; such a graph contains the monotonicity information about f.

Information about convexity of the graph is found by computing f'' and determining the sign: When $f'' > 0$, the graph is convex, and when $f'' < 0$ the graph is concave. Points where the convexity of a graph changes are geometrically interesting:

Definition 9.19 Let f be continuous at x_0 and twice-differentiable on a deleted neighborhood of x_0. The point $\big(x_0, f(x_0)\big)$ is an *inflection point* or a *flex*[2] if

- $\lim(f', x_0^-) = \lim(f', x_0^+)$ (possibly $+\infty$ or $-\infty$),
- f'' changes sign at x_0, and f'' is non-vanishing on some deleted interval about x_0.

Geometrically, the graph has a tangent line at each point in a neighborhood of x_0 and the convexity changes at x_0, so the graph is "S-shaped". By Darboux' theorem, an inflection point corresponds to a zero of f''

[2]In mathematics, "flex" is a noun.

or a point at which f'' does not exist. Not every such point is a flex, however.

Using the critical points as scaffolding and fleshing out the graph using convexity information generally gives an accurate graph. If more quantitative information is needed, plot a few points by direct computation. If the equation $f(x) = 0$ can be solved exactly, those points should be plotted.

Example 9.20 Suppose we wish to sketch the graph of $f(x) = x^{2/3}(x^3 - 1)$. First we multiply out and differentiate:

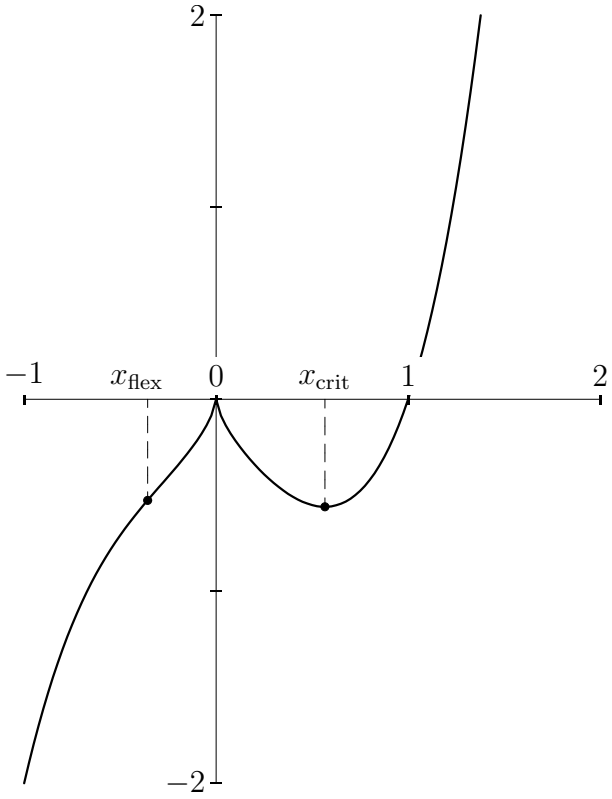
$$f'(x) = \tfrac{11}{3}x^{8/3} - \tfrac{2}{3}x^{-1/3} = \tfrac{1}{3}x^{-1/3}(11x^3 - 2)$$
$$f''(x) = \tfrac{88}{9}x^{5/3} + \tfrac{2}{9}x^{-4/3} = \tfrac{2}{9}x^{-4/3}(44x^3 + 2)$$

Note that it is easier to differentiate after expanding, but easier to find critical points after factoring. In practice, you will probably want to compute expanded and factors forms of the first two derivatives.

There is one real critical point, $x_{\text{crit}} = \sqrt[3]{2/11}$, and $f'(0)$ does not exist. Since the sign of $\sqrt[3]{x}$ is the same as the sign of x while $(11x - 2) < 0$ near 0, we find that $\lim(f', 0^-) = +\infty$ and $\lim(f', 0^+) = -\infty$. Consequently, $f' > 0$ if $x < 0$ or if $x > \sqrt[3]{2/11}$, and $f' < 0$ if $0 < x < \sqrt[3]{2/11}$.

The second derivative vanishes at $x_{\text{flex}} = -\sqrt[3]{1/22}$ and is undefined at 0. Since $x^{4/3} \geq 0$ for all real x, we see that $\lim(f'', 0) = +\infty$. There is, consequently, one inflection point, since f'' changes sign where the term in parentheses is 0. The graph is concave for $x < -\sqrt[3]{1/22}$ and convex for $x > -\sqrt[3]{1/22}$.

The equation $f(x) = 0$ has two real solutions, $x = 0$ and $x = 1$. We plot these points, as well as the critical point and the flex, then sketch the curve using our monotonicity and convexity information.



The picture above shows the graph at true scale. □

Acceleration

If a function represents position as a function of time, then the first derivative represents the physical concept of velocity, and the second derivative represents *acceleration*.

It is a remarkable physical fact that acceleration can be measured via "local experiments", those that do not involve looking at the rest of the universe. Imagine you are in an airplane, flying at constant speed and altitude. Aside from noise of the engines, you have no physical evidence you are "in motion". If you throw a ball across the aisle, it appears to move exactly as if the plane were sitting on the ground. Liquids poured from a can will fall into a cup in the expected way. Your weight as measured by a scale is the same as on the ground. By contrast, if the plane suddenly dives, climbs, or turns sharply, observable changes will occur: the ball will follow a strange path, possibly curving to one side; the drink may miss the cup, or (if the plane goes into "free-fall")

form a ball and hang motionless; your weight may increase or decrease as you stand on the scale. These effects are substantial for fighter pilots and astronauts, and in extreme cases can result in blackout from blood pooling in the legs and feet.

Airlines recommend that passengers keep their seat belts fastened during the entire flight, because there have been rare accidents in which a plane dived sharply for a fraction of a second, briefly "reversing gravity" in the cabin, and causing unsecured passengers to fall upward out of their seats and be seriously injured when they "landed" on the baggage compartments overhead. If you have flown you have almost certainly experienced the sensation of "turbulence" or an "air pocket". Your inner ear is an extremely sensitive gauge of equilibrium; when the plane is flying straight, level, and at constant speed, your inner ear cannot tell you are in motion, but if the plane accelerates, your inner ear will register the change, usually as a sensation of falling or nausea. NASA uses a specially modified Boeing 707 to train astronauts in a nearly weightless environment. The plane (nicknamed the "Vomit Comet" for obvious reasons) climbs to an altitude of about 40,000 feet, then enters a parabolic arc that matches the trajectory of a freely-falling body. This path is followed for about 30 seconds before the pilot must pull up the nose of the plane, which (incidentally) causes the plane to experience "greater than normal gravity" for a few seconds.

As recently as the early 1800s, some scientists believed that traveling at speeds in excess of about 15 mph would result in serious physical harm. Steam locomotives thoroughly debunked this idea, and we now know that velocity by itself is not merely harmless, but *physically meaningless* in an absolute sense. Acceleration, by contrast, *does* have absolute meaning, and manifests itself as a force (in the sense of physics). The fact that "acceleration can be felt but motion at constant speed cannot" is the basis of the adage, "It's not the fall that kills you, it's the sudden stop at the end."

9.5 Indeterminate Limits

There is a powerful calculational tool, l'Hôpital's rule, that harnesses the machinery of derivatives to the task of evaluating indeterminate limits. As is the case for many results of this chapter, there is a compelling Leibniz notation interpretation for the result, but the actual proof depends on a seemingly unrelated technical result.

The Cauchy Mean Value Theorem

Theorem 9.21. *Let f, $g : [a, b] \to \mathbf{R}$ be continuous, and differentiable on (a, b). There exists a point $x_0 \in (a, b)$ such that*

$$f'(x_0)\big(g(b) - g(a)\big) = g'(x_0)\big(f(b) - f(a)\big).$$

Proof. The ordinary mean value theorem is the special case $g = \mathrm{id}$, the identity function, and the Cauchy version is proven with an analogous trick. Define $h : [a, b] \to \mathbf{R}$ by

$$h(x) = f(x)\big(g(b) - g(a)\big) - g(x)\big(f(b) - f(a)\big).$$

It is immediate that h satisfies the hypotheses of Rolle's theorem, so there exists an $x_0 \in (a, b)$ with $h'(x_0) = 0$, as claimed. □

L'Hôpital's Rule

The corollary of the Cauchy mean value theorem that is known to calculus students as *l'Hôpital's rule* was in fact proven by John Bernoulli. The Marquis de l'Hôpital was a wealthy patron of Bernoulli but a mediocre mathematician. It is sometimes joked that l'Hôpital's rule is the best theorem that money can buy.

Theorem 9.22. *Suppose f and g are differentiable in some deleted neighborhood of c, that $\lim(f, c) = \lim(g, c) = 0$, that g' is non-vanishing in some deleted interval about c, and that $\lim(f'/g', c)$ exists and is equal to ℓ. Then $\lim(f/g, c)$ exists and is equal to ℓ.*

In English: When attempting to evaluate a limit of a quotient, if the answer is formally 0/0, then differentiate the numerator and denominator (*do not* confuse this with the quotient rule!) and try to evaluate again. If the limit is ℓ, then the original limit was also ℓ.

Proof. By assumption, f'/g' is defined on some deleted interval about c, so there exists a $\delta > 0$ such that if $0 < |x - c| < \delta$, then $f'(x)$ and $g'(x)$ exist and $g'(x) \neq 0$. If necessary, re-define f and g to be 0 at c, so that f and g are continuous on $(c - \delta, c + \delta)$.

In the deleted δ-interval about c, the denominator g is non-vanishing, since otherwise g' would be zero somewhere Rolle's theorem. Thus f/g is defined on the deleted interval $N_\delta^\times(c)$. By the Cauchy mean value theorem, there exists a point $x_0 \in (c, c + \delta)$ such that

$$f'(x_0)\big(g(c + \delta) - g(c)\big) = g'(x_0)\big(f(c + \delta) - f(c)\big),$$

or, since $f(c) = g(c) = 0$ and $g(c+\delta) \neq 0$,

$$\frac{f(c+\delta)}{g(c+\delta)} = \frac{f'(x_0)}{g'(x_0)}.$$

Taking the limit as $\delta \to 0$ completes the proof, since $x_0 \to c$ as $\delta \to 0$. □

It is very important not to apply the formalism without checking the hypotheses; in English, do not attempt to apply l'Hôpital's rule if the limit is not formally $0/0$. To see why, consider the mistaken calculation

$$\lim_{x\to 0} \frac{1}{x^2} \overset{\text{oops!}}{=} \lim_{x\to 0} \frac{0}{2x} = \lim_{x\to 0} \frac{0}{2} = 0.$$

In addition, the *converse* of l'Hôpital's rule is false. If $\lim(f'/g', c)$ fails to exist, then no information is gained; the original limit may or may not exist. It is still possible to say something, though, see Exercise 9.24.

Example 9.23 Let n be a positive integer. By l'Hôpital's rule,

$$\lim_{x\to 1} \frac{x^n - 1}{x - 1} = \lim_{x\to 1} \frac{nx^{n-1}}{1} = n,$$

in accord with the geometric sum formula, Exercise 2.16. By Exercise 9.4, the conclusion extends to arbitrary *rational* r. □

L'Hôpital's rule may be applied repeatedly in the event $f'(c) = g'(c) = 0$. If, for some positive integer k, the limit $\lim(f^{(k)}/g^{(k)}, c)$ exists and is equal to ℓ (and *all* previous applications of l'Hôpital's rule have given $0/0$), then the original limit exists and is equal to ℓ.

Example 9.24 A single application of l'Hôpital's rule gives us

$$\lim_{x\to 0} \frac{e^x - x - 1}{x^2} = \lim_{x\to 0} \frac{e^x - 1}{2x}$$

since exp is its own derivative. The latter is still formally $0/0$, so we apply l'Hôpital again. The resulting expression can be evaluated by setting $x = 0$, and we find that the limit is $1/2$. □

L'Hôpital's Rule at $+\infty$

L'Hôpital's rule has a version for limits at $+\infty$ that will be used repeatedly.

Theorem 9.25. *Let f and g be differentiable functions. Assume that g' is non-vanishing on some interval $(R, +\infty)$, and that*

$$\lim(f, +\infty) = 0 = \lim(g, +\infty).$$

If $\lim(f'/g', +\infty) = L$, *then* $\lim(f/g, +\infty) = L$.

Note that the conclusion is that the quotient f/g *has a limit* at $+\infty$, and that this limit is the same as the limit of f'/g'. Formally, this is the same as l'Hôpital's rule at finite points.

Proof. Because g' is non-vanishing on some interval $(R, +\infty)$, g is monotone on $(R, +\infty)$. Since $\lim(g, +\infty) = 0$, it follows that g itself is non-vanishing on $(R, +\infty)$, so the quotient f/g is defined on this interval. Because f and g have finite limit at $+\infty$, we may assume f and g are bounded on $(R, +\infty)$.

An argument based on the Cauchy mean value theorem is possible, but slightly involved, because the hypotheses of the theorem are not amenable to using closed, bounded intervals. Instead, we use the "change of variable" $y = 1/x^2$. This choice is dictated by the wish that x lie in a deleted interval about 0 iff y is in a deleted interval about $+\infty$. Define

$$F(x) = f(1/x^2), \quad G(x) = g(1/x^2) \qquad \text{for } x \neq 0.$$

You can check that F and G satisfy the hypotheses of Theorem 9.22. By the chain rule, we have $F'(x) = -(2/x^3)f'(1/x^2)$ and similarly for g, so

$$\frac{F'(x)}{G'(x)} = \frac{f'(1/x^2)}{g'(1/x^2)}.$$

Theorem 9.22 implies $\lim(f/g, +\infty) = \lim(F/G, 0) = L$, as claimed. □

Exercises

Exercise 9.1 Consider the absolute value function $f(x) = |x|$ on the interval $[-1, 1]$. Does there exist an $x_0 \in (-1, 1)$ such that

$$f'(x_0) = \frac{f(1) - f(-1)}{1 - (-1)} = 0?$$

Why does this not contradict the mean value theorem? Answer the same question for the function $g(x) = x/|x|$ defined for $x \neq 0$. ◇

Exercise 9.2 Let $f(x) = 1/x$ for $x \neq 0$. Show that $f'(x) < 0$ for all x in the domain of f. Is f a decreasing function? Is f decreasing when restricted to an interval in its domain? Explain. ◇

Exercise 9.3 Let $f(x) = x^3$ for $x \in \mathbf{R}$. Show that f has a critical point. Is f increasing on $\mathbf{R}$? ◇

Exercise 9.4 Use the result of Example 9.13 to show that if $r = p/q$ is rational, then $\frac{d}{dx}x^r = rx^{r-1}$. ◇

Exercise 9.5 Using the chain rule and Exercise 9.4, find the derivatives of the following, and sketch the graphs. Be sure to give the domain of each function and the domain of the derivative.

(a) $f(x) = \sqrt{1 - x^2}$

(b) $f(x) = 1/\sqrt{1 - x^2}$

(c) $f(x) = \sqrt{1 + x^2}$

(d) $f(x) = x/\sqrt{1 + x^2}$

It may be helpful to find the vertical and/or horizontal asymptotes, as appropriate. ◇

Exercise 9.6 Let $f : [-2, 2] \to \mathbf{R}$ be defined by

$$f(x) = \begin{cases} (1 - x^2)^2 & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Prove that f is differentiable (pay close attention to the points $x = \pm 1$), and sketch the graphs of f and f' on the same set of axes. ◇

Exercise 9.7 Let $\psi : \mathbf{R} \to \mathbf{R}$ be the pseudo-sine function of Example 9.11.

(a) Show that ψ is odd, express the derivative in terms of the Charlie Brown function, and find the critical points and extrema.

(b) Sketch the graph of ψ on $[-3, 3]$. Note that (9.1) holds only on $[-1, 1]$.

◇

Exercise 9.8 Let $f : \mathbf{R} \to \mathbf{R}$ be $\mathscr{C}^1$, and assume f' is non-constant and periodic with period 1.

(a) Prove that there exists a real c such that $f(x+1) = f(x) + c$ for all $x \in \mathbf{R}$.

(b) Prove that $g(x) = f(x) - cx$ is periodic with period 1.

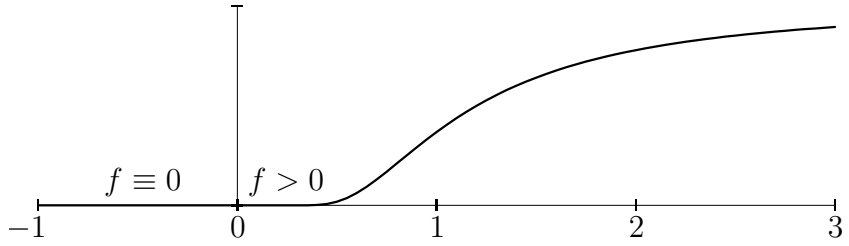
(c) Give an example of a non-periodic function whose derivative is non-constant and periodic, and sketch the graph of f.

◇

Exercise 9.9 Sketch the locus Z of the equation $y^2 = x^2 + x^3$, and show that there exists a continuous function $f : [-1, +\infty) \to \mathbf{R}$, differentiable on $(-1, +\infty)$, such that Z is the union of the graphs of f and $-f$. (You will need to patch two functions at 0.) ◇

Exercise 9.10 Let $f : \mathbf{R} \to \mathbf{R}$ be defined by

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$



(a) Use the previous exercise to show that if p is a polynomial, then

$$\lim_{x \to 0} p\left(\tfrac{1}{x}\right) f(x) = 0.$$

(b) Use the chain rule to compute $f'(x)$ if $x > 0$, and evaluate $\lim(f', 0)$ with full justification. Use your answer to prove that f is differentiable on $\mathbf{R}$.

(c) Use induction on n to prove that f is n times differentiable, even at 0.
Suggestion: Prove inductively that if $x > 0$, then $f^{(k)}(x) = p_k\left(\tfrac{1}{x}\right) f(x)$ for some polynomial p_k.

Thus f is smooth ($\mathscr{C}^\infty$). ◇

Exercise 9.11 In Exercise 8.3, you showed that

$$\sum_{k=1}^{n} kx^k = \frac{nx^{n+2} - (n+1)x^{n+1} + x}{(x-1)^2}$$

$$\sum_{k=1}^{n} k^2x^k = \frac{n^2x^{n+2} - (2n^2+2n-1)x^{n+1} + (n+1)^2x^n - (x+1)}{(x-1)^3}$$

for $x \neq 1$. Use l'Hôpital's rule at $x = 1$ to evaluate the sums on the left at $x = 1$. ◇

Convexity

Exercise 9.12 Let R_1 and R_2 be convex sets in the plane. Prove that $R_1 \cap R_2$ is convex. ◇

Exercise 9.13 Let a and b be positive. The set

$$(*) \qquad \{(x,y) \mid \tfrac{x^2}{a^2} + \tfrac{y^2}{b^2} \leq 1\}$$

is an ellipse, cf. Figure 9.5.

(a) Solve for y as a function of x to express the boundary of the ellipse as a union of graphs.

(b) Prove that the function whose graph is the top half of the ellipse is concave, and that the bottom half is convex.

(c) Prove that the ellipse $(*)$ is a convex set in the plane.

Exercise 9.12 should be helpful. ◇

Exercise 9.14 Let $I \subset \mathbf{R}$ be an interval. Show that f is convex on I iff

$$f\big((1-t)a + tb\big) \leq (1-t)f(a) + tf(b), \qquad 0 \leq t \leq 1$$

for all a and b in I. ◇

Exercise 9.15 Prove that the indicator function $\chi_{\{0\}} : [0,1] \to \mathbf{R}$ is convex.
Hint: You can find the secants explicitly. ◇

Exercise 9.16 Prove that if $f : [0,1] \to \mathbf{R}$ has a jump discontinuity in $(0,1)$, then f is not convex. ◇

Exercise 9.17 Prove that if f is twice-differentiable and $f'' \neq 0$ on $[a,b]$, then f vanishes at most twice in $[a,b]$.
Hint: Use the mean value theorem directly; do not assume f is $\mathscr{C}^2$.

The intuitive principle is that if f'' is non-vanishing (nowhere zero), then f vanishes at most twice on each interval contained in the domain of f. If the domain of f is not an interval, then f may vanish more than twice. ◇

Exercise 9.18 Let $f : [0,+\infty) \to \mathbf{R}$ be a non-increasing, convex function. Can f have a jump discontinuity? Two jump discontinuities? What if we do not assume f is non-increasing? As always, give proof or counterexamples for your claims. ◇

Exercise 9.19 This exercise establishes *Hölder's inequality*: If $\frac{1}{p}+\frac{1}{q} = 1$ and if f and g are integrable on $[a,b]$, then

$$(**) \qquad \Big|\int_a^b fg\Big| \leq \int_a^b |fg| \leq \Big(\int_a^b |f|\Big)^{1/p}\Big(\int_a^b |g|\Big)^{1/q}$$

(Recall that fg is integrable by Exercise 7.12.)

(a) Let $\alpha \in (0,1)$. Prove that $t^\alpha \leq \alpha t + (1-\alpha)$ for all $t \geq 0$.

(b) Let $\beta = 1-\alpha$, and note that $\beta \in (0,1)$. Prove that

$$u^\alpha v^\beta \leq \alpha u + \beta v \qquad \text{for all } u,\, v > 0.$$

Suggestion: Set $t = u/v$ in part (a).

(c) Let $p > 1$, and set $q = p/(p-1)$, so that $\frac{1}{p}+\frac{1}{q} = 1$. Show that

$$AB \leq \tfrac{1}{p}A^p + \tfrac{1}{q}B^q \qquad \text{for all } A,\, B \geq 0.$$

Suggestion: Use part (b) with appropriate changes of variable.

(d) (Hölder's inequality for finite sequences) Show that if p and q are as in (c), and if a_k and b_k are real numbers for $1 \leq k \leq n$, then

$$\sum_{k=1}^n |a_k b_k| \leq \Big(\sum_{k=1}^n |a_k|^p\Big)^{1/p}\Big(\sum_{k=1}^n |b_k|^q\Big)^{1/q}.$$

Hint: Set $A_k = a_k/(\sum_k |a_k|^p)^{1/p}$, etc., and use part (c).

(e) Prove equation (∗∗). There are a couple of ways to proceed; either use part (d) and Riemann sums, or part (c) with A and B suitable integrals.

Hölder's inequality is of great technical importance in analysis. The special case $p = q = 2$ is the *Schwarz inequality.* ◇

L'Hôpital's Rule

Exercise 9.20 Use l'Hôpital's rule to calculate the following limits:

$$\lim_{x\to 1} \frac{\log x}{x-1} \qquad \lim_{x\to 0^+} x \log x \qquad \lim_{x\to+\infty} \frac{\log x}{x} \qquad \lim_{x\to+\infty} \frac{x}{\exp x}.$$

Suggestion for the second limit: write $x = 1/(1/x)$. ◇

Exercise 9.21 Use the results of the preceding exercise, and other techniques as needed, to evaluate the following limits:

$$\lim_{x\to+\infty} \frac{x^n}{\exp x} \quad \text{and} \quad \lim_{x\to 0} \frac{e^{-1/x^2}}{x^n} \quad \text{for } n \in \mathbf{N}; \qquad \lim_{x\to 0^+} x^x.$$

Despite the last limit, the expression 0^0 is undefined. ◇

Exercise 9.22 Let f be differentiable in a neighborhood of x. Evaluate

$$\lim_{h\to 0} \frac{f(x+h) - f(x-h)}{2h}.$$

Find a discontinuous function for which this limit exists. ◇

Exercise 9.23 Let f be twice-differentiable in a neighborhood x. Evaluate

$$\lim_{h\to 0} \frac{f(x+h) + f(x-h) - 2f(x)}{h^2}.$$

Why do we not use this limit to define f''? ◇

Exercise 9.24 Suppose f and g are differentiable on $(c-\delta, c+\delta)$, that $\lim_{x\to c} f(x) = \lim_{x\to c} g(x) = 0$, and that $\lim_{x\to c} g'(x) = 0$ but $\lim_{x\to c} f'(x) = \ell \neq 0$. Prove that $\lim_{x\to c} f(x)/g(x)$ does not exist. ◇

Chapter 10

The Fundamental Theorems

We have introduced an operation (integration) that "adds up infinitely many infinitesimals", and another (differentiation) that "zooms in with factor infinity". Aside from technical issues, these operations are linear mappings $f \mapsto I_a f$ and $f \mapsto Df$, defined by

$$I_a f(x) = \int_a^x f(t)\,dt, \qquad Df(x) = f'(x).$$

We have argued informally that these operations should be inverse to each other, and have rigorously established some partial results in this direction. With the mean value theorem available, we are now ready to make a systematic and detailed study of the relationship between integration and differentiation.

10.1 Integration and Differentiation

Theorem 10.1. *Let $f : [a, b] \to \mathbf{R}$ be integrable, and let $F : [a, b] \to \mathbf{R}$ be the definite integral of f, defined by*

$$F(x) = \int_a^x f(t)\,dt \qquad \text{for } x \in [a, b].$$

If f is continuous at $c \in (a, b)$, then F is differentiable at c, and $F'(c) = f(c)$.

Theorem 10.1 is often called the *First Fundamental Theorem of Calculus*, a name that emphasizes its central role in the calculus. In Leibniz notation, the conclusion is written

$$\frac{d}{dx}\int_a^x f(t)\,dt = f(x),$$

which emphasizes the inverse nature of integration and differentiation.

As usual with a complicated theorem, it is tempting to memorize only the conclusion. You are cautioned that the conclusion is not generally true if f is discontinuous: The derivative of the integral may fail to exist at all, and may have the "wrong" value even if it does exist.

Proof. Fix $c \in (a, b)$. By hypothesis, $f = f(c) + o(1)$ at c, so for each $\varepsilon > 0$ there exists an open δ-interval contained in (a, b) on which $f = f(c) + A(\varepsilon)$. If $0 < |h| < \delta$ then the Newton quotient $\Delta_c F(h)$ is defined, and is equal to

$$\Delta_c F(h) = \frac{1}{h}\Big(F(c+h) - F(c)\Big) = \frac{1}{h}\int_c^{c+h} f(t)\,dt.$$

As noted in Chapter 7, this is the average value of f on the interval with endpoints c and $c + h$ (even if $h < 0$). By Theorem 7.20,

$$\Delta_c F(h) = \frac{1}{h}\big(f(c)h + A(\varepsilon)\,h\big) = f(c) + A(\varepsilon)$$

on the open δ-interval about c. Since $\varepsilon > 0$ was arbitrary, we have shown that $\Delta_c F(h) = f(c) + o(1)$ near $h = 0$, or $F'(c) = f(c)$. □

The first fundamental theorem says what happens when a *continuous* function f is integrated and the integral differentiated: The function f is recovered! The second fundamental theorem of calculus treats the opposite question, in which a derivative is integrated.

Theorem 10.2. *If $f : [a, b] \to \mathbf{R}$ is integrable, and if $f = F'$ for some function F, then*

$$\int_a^b f = F(b) - F(a).$$

Proof. Let $P = \{t_i\}_{i-0}^n$ be a partition of $[a, b]$. By the mean value theorem applied to F, for each $i = 1, \ldots, n$ there exists a point $x_i \in (t_{i-1}, t_i)$ such that

$$F(t_i) - F(t_{i-1}) = F'(x_i)(t_i - t_{i-1}) = f(x_i)\Delta t_i.$$

If m_i and M_i are the inf and sup of f on the ith subinterval, then

$$m_i \Delta t_i \le F(t_i) - F(t_{i-1}) \le M_i \Delta t_i \qquad \text{for } i = 1, \ldots, n.$$

Summing over i shows that

$$L(f,P) \le F(b) - F(a) \le U(f,P) \qquad \text{for every partition } P.$$

Since f is integrable and the middle term does not depend on the choice of partition, the value of the integral must be $F(b) - F(a)$. □

The proof of Theorem 10.2 is a rigorous version of the intuitive argument given in Chapter 6, in which integration of the differential

$$F'(t)\,dt = \frac{dF}{dt}dt = dF = F(t+dt) - F(t)$$

gives rise to a formally telescoping sum. In Leibniz notation, the conclusion of the second fundamental theorem reads

$$\int_a^x dF = \int_a^x \frac{dF}{dt}\,dt = F(x) - F(a).$$

As the notation suggests, we usually regard a as fixed and x as variable. However, either a or x can be regarded as a "variable" in Theorem 10.3, since both are arbitrary. Again, you should not memorize the conclusion and forget the hypotheses; it is essential to assume that F' is integrable.

This is a good time to recall the notation

$$F(x) - F(a) = F\Big|_a^x = F(t)\Big|_{t=a}^{t=x} = \Big[F(t)\Big]_{t=a}^{t=x}.$$

It is common to see "$F(t)\big|_a^x$" as well, though strictly speaking this is bad syntax.

The difference $F\big|_a^x$ may be regarded as a "0-dimensional integral" (a.k.a. "sum of function values, counted with orientation") of F over the boundary of the interval $[a,x]$. The fundamental theorem asserts this is the same as the "1-dimensional integral" of the differential $dF = F'(t)\,dt$ over $[a,x]$. The proper setting for these remarks is the calculus of several variables.

Since a continuous function is integrable on every closed interval, Theorem 10.2 has an immediate corollary:

Theorem 10.3. *If $F : (\alpha,\beta) \to \mathbf{R}$ is $\mathscr{C}^1$, then*

$$\int_a^x F'(t)\,dt = F(x) - F(a) \qquad \textit{for all } a,\ x \in (\alpha,\beta).$$

Integration *vs.* Antidifferentiation

An *antiderivative* of f is a function F with $F' = f$. Theorem 10.3 says that integrating a *continuous* function f over the interval $[a, x]$ is tantamount to finding an antiderivative of f and computing a difference of function values. Note the powerful implication: a *single* antiderivative allows us to evaluate an integral for *all* x *in some interval*. Finding an antiderivative is often considerably easier than computing lower sums and taking the supremum, even for a *single* value of x.

The importance of the fundamental theorem is twofold:

- (Practical) It greatly simplifies calculation of integrals of functions for which an antiderivative can be found explicitly.
- (Theoretical) It exhibits an antiderivative of a continuous function whether or not an antiderivative can be found by other means.

The practical importance alone is complete justification of the fundamental theorem in most calculus courses. However, the theoretical importance should not be overlooked: We have already seen, in Exercise 7.17, how the properties of a function defined as a definite integral can be studied (that is, how interesting functions can be defined as integrals).

Theorem 10.2 describes the close link between integration and antidifferentiation, and many calculus texts (as well as working scientists and mathematicians) use the integral sign to *denote* antiderivatives, as in

$$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C,$$

the ubiquitous "$+C$" representing an arbitrary additive constant.[1] Many students are left with the impression that integration *is* antidifferentiation, perhaps even by definition. The reason is probably human psychology: The definition of integration is relatively complicated (partitions, sums, and suprema), antiderivatives are simple by comparison, and "most of the time" they're functionally equivalent. However, integration and antidifferentiation are not the same thing, and therein lies the miracle of differential and integral calculus. An antiderivative is often easy to find, but has no obviously interesting interpretation.

[1] We avoid such notation; "x" is a dummy variable on the left but not on the right.

An integral is rich with meaning (total change), but is laborious to compute from the definition.

> *The fundamental theorem of calculus gives a method of working easily with quantities of great theoretical and practical interest.*

A more subtle point is that Theorems 10.1 and 10.3 are stated for *continuous* integrands. It is simply *not true* that integration and differentiation are inverse operations, even up to the additive constant. Precisely,

- There exist (discontinuous) integrable functions f whose integral is everywhere differentiable, but such that

$$f(x) \neq \frac{d}{dx} \int_a^x f(t)\, dt$$

  for infinitely many x. The denominator function of Example 3.11 has this property, see Exercise 7.9.

- There exist differentiable functions F (with discontinuous derivative) such that F' is not integrable, so Theorem 10.2 is not even applicable.

Despite these cautions, the fundamental theorems can be strengthened; for example, the conclusion of Theorem 10.3 remains true for differentiable functions F whose derivatives are unbounded at finitely many points if we enlarge the definition of "integral" to include improper integrals, see Exercise 10.20.

10.2 Antidifferentiation

Because integration and antidifferentiation are so closely related for continuous functions, every calculational theorem about derivatives (the chain rule and product rule especially) corresponds to a useful tool for computing integrals. The *method of substitution* arises from the chain rule, while the technique of *integration by parts* corresponds to the Leibniz rule.

Strikingly (given the state of affairs with derivatives), *there is no general formula or algorithm for finding antiderivatives of products, quotients, and compositions.* This is not an assertion of mathematical

ignorance, but a fact of life. The best one can do is follow the dictum,[2] "Shoot first, and call whatever you hit 'the target'." Specifically, the strategy is to compile a table of known derivatives; entries of this table are functions that we know how to antidifferentiate.

The vague principle "antidifferentiation is non-algorithmic" summarizes the lessons of several theorems. This issues is studied in greater depth in Chapter 15, but already we can give an indication of what goes wrong. Recall that a "rational function" is a quotient of polynomials, and that an "algebraic function" is defined implicitly by a polynomial in two variables. Every rational function is algebraic; formally, if $f(x) = p(x)/q(x)$ with p and q polynomial, then $F(x, y) = p(x) - yq(x)$ defines f implicitly. Theorems of Chapter 8 imply that the derivative of a rational function is rational, and that the derivative of an algebraic function is algebraic.

The natural logarithm, an antiderivative of the (rational) reciprocal function, is not even algebraic. This is not an isolated example, but a feature of a "randomly chosen" rational function. Further, we will see in Chapter 13 that the inverse trigonometric functions, none of which is algebraic, all have algebraic derivatives, and that arctan even has rational derivative. As for rational functions, a "generic" algebraic function does not have an algebraic antiderivative.

Despite these dire cautions, a great many functions can be antidifferentiated explicitly. However, the flavor of the subject is more that of an art than a science. At this point in the book, the State of the Art is not very impressive:

Proposition 10.4. *Let $r \neq -1$ be a rational number, and define $f(x) = x^r$ for $x > 0$. The power function $F(x) = x^{r+1}/(r+1)$ is an antiderivative of f. The natural logarithm is an antiderivative of the reciprocal function $f(x) = x^{-1}$.*

Sums of constant multiples of these rational-power functions (for example, polynomials) are trivially handled as well. The fundamental theorem allows us to compute, for example, that

$$\int_0^1 (t + 2\sqrt{t})\, dt = \left(\frac{t^2}{2} + 2\frac{t^{3/2}}{3/2} \right) \Bigg|_{t=0}^{t=1} = \frac{1}{2} + \frac{4}{3}.$$

However, we are currently stymied by integrands such as $\sqrt{1+t^2}$, $t\sqrt{1+t^2}$, and $\sqrt{1+t^{1/2}}$. Of course, these functions *have* antideriva-

[2] Due to Ashleigh Brilliant, used with permission.

tives (namely their definite integrals on suitable intervals), but we do not yet know how to write these antiderivatives in "closed form".

Substitution

Just as the chain rule allows differentiation of composite functions, the "method of substitution" or "change of variables theorem" allows antidifferentiation of suitable functions.

Theorem 10.5. *Let $g : (\alpha, \beta) \to \mathbf{R}$ be $\mathscr{C}^1$, $[a, b] \subset (\alpha, \beta)$, and assume f is a continuous function whose domain contains $g([a, b])$. Then*

$$\int_a^b (f \circ g)\, g' = \int_{g(a)}^{g(b)} f.$$

Proof. Let F be an antiderivative of f, and set $G = F \circ g : [a, b] \to \mathbf{R}$. The function G is continuously differentiable, and the chain rule implies $G' = (f \circ g)\, g'$. By the second fundamental theorem of calculus,

$$\begin{aligned} \int_a^b (f \circ g)\, g' &= \int_a^b G' = G\Big|_a^b \\ &= F\big(g(b)\big) - F\big(g(a)\big) = F\Big|_{g(a)}^{g(b)} \\ &= \int_{g(a)}^{g(b)} F', \end{aligned}$$

and since $F' = f$ the theorem is proved. □

In traditional notation with dummy variables, the conclusion of the change of variables theorem is written

$$\int_a^b \big(f \circ g\big)(t)\, g'(t)\, dt = \int_{g(a)}^{g(b)} f(u)\, du. \tag{10.1}$$

This formulation is particularly compelling in Leibniz notation for derivatives. The "substitution" $u = g(t)$ is differentiated, yielding

$$du = g'(t)\, dt = \frac{du}{dt}\, dt,$$

which looks exactly like cancellation of fractions.[3] To change the limits on the integral, note that if $t = a$, then $u = g(a)$. Similarly $t = b$

[3]Remember that we have not assigned meaning to isolated infinitesimal expressions.

corresponds to $u = g(b)$. Formal substitution converts the left-hand size of (10.1) into the right-hand side. Again, this means Leibniz notation is well-chosen, not that Theorem 10.5 is a tautology.

Example 10.6 Consider

$$\int_0^2 (1 + t - t^3)^{10} (1 - 3t^2)\, dt.$$

Evaluating directly from the definition is hopeless, as it would first require multiplying out (to get a polynomial of degree 32), and then evaluating upper and lower sums and finding the infimum and supremum, respectively. However, the derivative of $(1 + t - t^3)$ with respect to t is $1 - 3t^2$. If we set $u = (1 + t - t^3)$, then $u = 1$ when $t = 0$ and $u = -5$ when $t = 2$, so

$$\int_0^2 (1+t-t^3)^{10} (1-3t^2)\, dt = \int_1^{-5} u^{10}\, du = \frac{u^{11}}{11}\bigg|_{u=1}^{u=-5} = \frac{1}{11}\Big((-5)^{11} - 1\Big).$$

If an antiderivative had been sought, it could have been found after antidifferentiating by setting $u = (1 + t - t^3)$ instead of plugging in numerical limits. In Leibniz notation,

$$\frac{d}{dt}\frac{1}{11}(1 + t - t^3)^{11} = (1 + t - t^3)^{10} (1 - 3t^2),$$

as is clear from the chain rule. □

Exercises

"Standard" calculus exercises include calculation of lots of antiderivatives. While this is an important skill, it is increasingly possible to rely on symbolic manipulation programs; techniques of integration may fall by the wayside for non-mathematicians. It *is* worthwhile to know when a given function can be expected to have an explicit antiderivative. In many cases, algebra will convert an apparently impossible function into one that is easily integrated.

Exercise 10.1 The following functions are given by their value at t. Find an antiderivative, and check your answer by differentiation. Some

of them can be done more than one way.

(a_1)	$(1+t)^3$	(a_2)	$t(1+t^2)^3$	(a_3)	$(1+t^2)^3$
(b_1)	$\sqrt{1+t}$	(b_2)	$2t\sqrt{1+t^2}$	(b_3)	$t/\sqrt{2+t^2}$
(c_1)	$(1+\sqrt{t})^2$	(c_2)	$(t+t^{-1})^3$	(c_3)	$(1-t^{-2})(t+t^{-1})^3$
(d_1)	$(t+\sqrt{t})^2$	(d_2)	$(\sqrt{t}+1/\sqrt{t})^2$	(d_3)	$(t^{1/2}+t^{3/2})^3(1+3t)/\sqrt{t}$
(e_1)	$1/(1-t^2)$	(e_2)	$t/(1-t^2)$	(e_3)	$1/\big((t-1)^2(1+t)\big)$

Note that you may need different techniques even for a single part. ◇

Exercise 10.2 Evaluate $\displaystyle\int_0^x \frac{(1+t)}{(2+2t+t^2)^3}\,dt$ for $x \in \mathbf{R}$ ◇

Exercise 10.3 Suppose $f : \mathbf{R} \to \mathbf{R}$ is continuous.

(a) Define $G : \mathbf{R} \to \mathbf{R}$ by

$$G(x) = \int_0^{x^2} f = \int_0^{x^2} f(t)\,dt.$$

Prove that G is differentiable, and find $G'(x)$.
Suggestion: Write G as the composition of two functions you understand well, and use the chain rule.

(b) Define $H : \mathbf{R} \to \mathbf{R}$ by

$$H(x) = \int_x^{x^2} f.$$

Show that H is differentiable, and find $H'(x)$.

(c) Suppose generally that ϕ and ψ are differentiable on (α, β), and define

$$\Phi(x) = \int_{\psi(x)}^{\phi(x)} f \qquad \text{for } x \in (\alpha, \beta).$$

Show that Φ is differentiable, and find $\Phi'(x)$ in terms of f, ϕ, and ψ.

◇

Exercise 10.4 Does there exist an integrable function $f : [0,1] \to \mathbf{R}$ such that

$$\int_0^x f = 1 \qquad \text{for all } x \in (0,1]?$$

Explain. ◇

Exercise 10.5 Does there exist an integrable function $f : [-1,1] \to \mathbf{R}$ such that

$$\int_{-1}^x f = \sqrt{1-x^2} \qquad \text{for all } x \in [-1,1]?$$

Explain. What if f is only required to be *improperly* integrable? ◇

Exercise 10.6 Suppose f and g are integrable functions on $[a,b]$, and that

$$\int_a^x f = \int_a^x g \qquad \text{for all } x \in [a,b].$$

Does it follow that $f(x) = g(x)$ for all x? What if f and g are continuous? ◇

Exercise 10.7 Let $u : \mathbf{R} \to \mathbf{R}$ be continuous. Prove that for each $c \in \mathbf{R}$, there is exactly one $\mathscr{C}^1$ function $f : \mathbf{R} \to \mathbf{R}$ satisfying the initial value problem

$$f' = u, \qquad f(0) = c.$$

◇

Exercise 10.8 Let $I \subset \mathbf{R}$ be an interval, and let $f : I \to \mathbf{R}$ be $\mathscr{C}^1$. Prove that there exist *non-decreasing* functions g_1 and g_2 with $f = g_1 - g_2$.
Hint: Consider the positive and negative parts of f'.

Let $f(x) = x - x^3/3$, $|x| \le 2$, see Figure 9.2. On the same set of axes, sketch f and a pair of non-decreasing functions whose difference is f. ◇

Exercise 10.9 Let $A : (-1,1) \to \mathbf{R}$ be the $\mathscr{C}^1$ function characterized by

$$A'(x) = \frac{1}{\sqrt{1-x^2}}, \qquad A(0) = 0.$$

(a) Prove that A is injective.

(b) By (a), A is invertible; let $S = A^{-1}$. Prove that $S' = \sqrt{1-S^2}$.

(c) Use (b) to show that S is $\mathscr{C}^2$, and fins S'' in terms of S.

◇

Exercise 10.10 Use the change of variables theorem to re-do Exercise 7.13. (Your calculation should be extremely brief. Good theoretical tools can be a tremendous labor-saving device.) ◇

Exercise 10.11 Let $f : \mathbf{R} \to \mathbf{R}$ be continuous.

(a) Define

$$g(x) = \int_0^x tf(t)\,dt, \qquad h(x) = \int_0^x xf(t)\,dt.$$

Find $g'(x)$ and $h'(x)$.
Suggestion: Rewrite h so that there is no "x" inside the integral.

(b) Use the result of part (a) to show that

$$\int_0^x f(u)(x-u)\,du = \int_0^x \left(\int_0^u f(t)\,dt\right) du.$$

Suggestion: Differentiate each side with respect to x. ◇

Exercise 10.12 Suppose u and v are continuously differentiable functions on $[a, b]$. Prove the *integration by parts formula*:

$$\int_a^b u\,v' = uv\Big|_a^b - \int_a^b v\,u',$$

and write it in Leibniz notation. Suggestion: Integrate the product rule for derivatives. ◇

Exercise 10.13 Formally "show" that $\displaystyle\int_{-1}^{1} \frac{1}{x^2}\,dx = -2$.

(a) The integrand is positive, but the integral is negative. What went wrong?

(b) Explain why, when you are asked to evaluate an improper integral, you must always prove separately that the integral exists.

The lesson is that formal calculation is fine when legitimate, but can lead to errors if applied mindlessly. ◇

Exercise 10.14 Use Exercise 10.12 to evaluate the following:

$$\int_1^b t^n \log t\,dt \qquad \int_0^1 t^n \log t\,dt.$$

Suggestion: Let $u(t) = \log t$ and $v(t) = t^n$. Note that the second integral is improper, so you must establish convergence. ◇

Exercise 10.15 Evaluate $\displaystyle\int_0^1 \frac{dt}{\sqrt{t}}$ and $\displaystyle\int_1^{+\infty} \frac{dt}{t^3}$. ◇

Exercise 10.16 This generalizes the preceding problem. Let $r > 0$ be rational; evaluate the improper integrals:

$$\int_0^1 \frac{dt}{t^r} \quad (0 < r < 1) \qquad \int_1^{+\infty} \frac{dt}{t^r} \quad (1 < r).$$

◇

Exercise 10.17 Evaluate the improper integral $\displaystyle\int_0^1 \frac{t}{\sqrt[3]{1-t^2}}\, dt$ ◇

Exercise 10.18 Determine which of the following improper integrals converge; you need not evaluate.

(a) $\displaystyle\int_0^{+\infty} \frac{t\, dt}{t^2+1}$

(b) $\displaystyle\int_{-1}^1 \frac{dt}{\sqrt{1-t^2}}$

(c) $\displaystyle\int_{-1}^1 \frac{t\, dt}{\sqrt{1-t^2}}$

Be sure to split the integral into pieces if necessary. ◇

Exercise 10.19 Let $\psi : \mathbf{R} \to \mathbf{R}$ be the pseudo-sine function, and consider the function

$$F(x) = \begin{cases} x^2\, \psi(1/x^2) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Prove that F is differentiable on $\mathbf{R}$, but that F' is unbounded near 0. Is it legitimate to write

$$F(x) = \int_0^x F'(t)\, dt$$

with the understanding that the integral is improper? ◇

Exercise 10.20 Suppose F is differentiable on $(\alpha, \beta) \supset [a, b]$, and that F' has only finitely many discontinuities. Prove that

$$\int_a^b F' = F(b) - F(a)$$

provided the left-hand side is interpreted as an improper integral. Exercise 10.19 gives an example of a function to which this result applies. ◇

Chapter 11

Sequences of Functions

At this stage we have developed the basic tools of the calculus, differentiation and integration, and have seen how they are related and how they can be used to study the behavior of functions. What we lack is a library of functions, especially logarithms, exponentials, and the trigonometric functions. (We have defined the natural logarithm, but have not yet proven identities like $\log(a^b) = b \log a$ and $e^{xy} = (e^x)^y$, so we are not yet in a position to manipulate logarithms and exponentials fluently.) There is a reason for our lack of knowledge: In order even to *define* non-algebraic functions precisely (in terms of axioms for $\mathbf{R}$), we must use non-algebraic operations, such as limits and suprema. If you are skeptical (and you should be!), try to define the cosine of a general angle without reference to geometry, or explain what is meant by $2^{\sqrt{2}}$, using only the axioms of $\mathbf{R}$.

One of the most concrete ways to incorporate limits into the definition of a function is via *power series*, namely "polynomials of infinite degree." An example is

$$f(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + \frac{x}{1} + \frac{x^2}{2 \cdot 1} + \frac{x^3}{3 \cdot 2 \cdot 1} + \cdots + \frac{x^k}{k!} \cdots . \tag{11.1}$$

Though literal addition of infinitely many terms is undefined, we may fix a specific x, regard (11.1) as a numerical series, and ask for which x the series converges. This is so-called "pointwise convergence." For technical reasons that will become apparent, it is instead better to regard the partial sums of (11.1) as a sequence of *functions*, and not merely to ask for which x the series converges, but to demand that the series converges "at the same rate" on some interval.

Before investigating convergence in detail, let's see what we stand to obtain from power series. Optimistically, if the right-hand side of (11.1) converges for all x in some interval, we hope to manipulate the resulting function as if it were a "polynomial with infinitely many terms"; specifically, we hope to differentiate and integrate term-by-term, to approximate the numerical value of the series by plugging in specific values of x, and so forth. Even more ambitiously, we might consider more general sequences of functions, and ask whether we can integrate or differentiate the limit function (which may be difficult) using the approximating functions (which in many cases is simple). There is no brief answer to these questions in general, but the situation for power series is as nice as possible: If a power series converges somewhere other than its center, then it converges on an interval, and on its interval of convergence may be manipulated exactly as if it were a polynomial. This, in one sentence, is the philosophical moral of this chapter.

Sequences and series are two ways of describing the same thing. Though series arise naturally in applications, we use sequence notation in deducing general theoretical results. Also, though our primary interest is power series, we study general sequences of functions first. This initial avoidance of unnecessary detail clarifies several technical issues.

11.1 Convergence

Let $I \subset \mathbf{R}$ be non-empty (usually an interval). Informally, a "sequence of functions" on I is an ordered collection $(f_n)_{n=0}^{\infty}$ of real-valued functions having domain I. Formally, a *sequence of functions* on I is a function

$$F : \mathbf{N} \times I \to \mathbf{R}, \qquad f_n = F(n, \ \cdot \ ) : I \to \mathbf{R}.$$

The geometric way a limit is taken is to regard the graph of f_n as a frame in an infinite movie (taken "at time n"), then to see what happens as the film runs on. Roughly, convergence of the sequence means the graphs "settle down" to an equilibrium graph as $n \to \infty$. It is not difficult to give a precise definition, but it takes lots of examples to shape intuition correctly, and a certain amount of hindsight to understand why the "best" definition of convergence is not the first one that comes to mind.

Pointwise Convergence

Let (f_n) be a sequence of functions on I. For each $x \in I$, the values $f_n(x)$ constitute a sequence of real numbers, and it makes sense to ask whether or not the limit exists for each x. If

$$f(x) := \lim_{n\to\infty} f_n(x) \tag{11.2}$$

exists for all $x \in I$, then the sequence (f_n) is said to *converge pointwise* to f, and the function f defined by equation (11.2) is the *pointwise limit* of the sequence (f_n).

Most students, if asked to define "convergence" of a sequence of functions, would probably choose a definition equivalent to pointwise convergence. However, as the following examples demonstrate, a pointwise limit may not possess desirable properties of the terms of its approximating sequence. We will consequently be led to seek a stronger criterion of convergence.

Example 11.1 Let $I = [0,1]$, and let $f_n : I \to \mathbf{R}$ be the piecewise linear function

$$f(x) = \begin{cases} 1 - nx & \text{if } 0 \le x \le 1/n \\ 0 & \text{if } 1/n < x \le 1 \end{cases}$$

The sequence (f_n) converges pointwise to a function f; clearly $f_n(0) = 1$ for all $n \in \mathbf{N}$, so $f(0) = 1$. If, instead, $x > 0$, then for every $n > 1/x$ we have $1/n < x$ and therefore $f_n(x) = 0$. It follows that $f(x) = 0$ for $x > 0$.

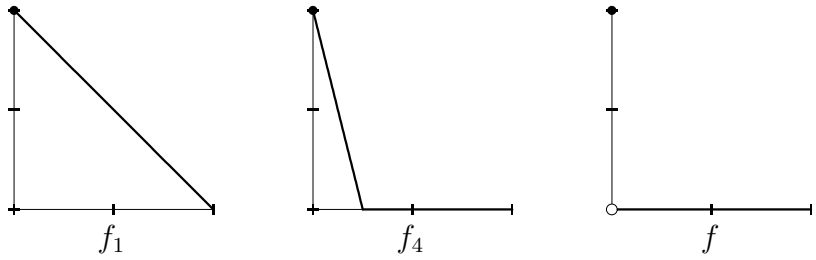


Figure 11.1: A sequence of continuous functions having discontinuous limit.

In summary, the sequence (f_n) converges pointwise to

$$f(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{if } x \in (0,1]. \end{cases}$$

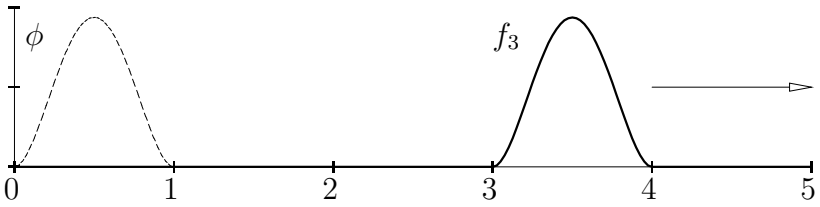


Figure 11.2: A bump disappearing at $+\infty$.

While each function f_n is continuous, the limit function is not. Passing to the limit of a pointwise convergent sequence can cause the graph to "break"; the graph of the limit is not the "limit of the graphs" in a naive geometric sense. □

Example 11.2 Consider the differentiable function $\phi : \mathbf{R} \to \mathbf{R}$ defined by

$$\phi(x) = \begin{cases} 30(x - x^2)^2 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and define $f_n(x) = \phi(x - n)$, Figure 11.2. The graph of ϕ has a "bump" of width 1 that encloses 1 unit of area, and the graph of f_n is translated to the right by n. In particular, f_n is identically zero on the interval $(-\infty, n]$ for each $n \in \mathbf{N}$.

If $x \in \mathbf{R}$, then there exists a natural number N such that $x < N$, which implies that $f_n(x) = 0$ for all $n > N$. Consequently,

$$f(x) := \lim_{n\to\infty} f_n(x) = 0 \qquad \text{for all } x \in \mathbf{R};$$

the sequence (f_n) converges pointwise to the zero function. □

Features of the f_n—in this case, a bump—are not generally inherited by the limit function. In this example, the bump moves to the right as n increases, and there is no upper bound on its distance to the origin. For every $x > 0$, no matter how large, the bump comes in like a wave from the left, then passes to the right of x. After that, nothing changes at x. In a sense, the bump "disappears at infinity."

Modifications of the preceding example may seem even more surprising. The sequence $(nf_n)_{n=1}^{\infty}$ converges pointwise to the zero function, though the bumps get arbitrarily large as they move to the right. Moreover, it is not necessary for the bumps to disappear "at infinity" in space:

Example 11.3 With ϕ as above, set $h_n(x) = n\phi(nx)$. It is left to you

(see Exercise 11.1) to show that the sequence (h_n) converges pointwise to the zero function, despite the fact that the graph of h_n has a "spike" of height n just to the right of 0 for each $n > 0$.

An additional feature of this example at first seems paradoxical: For each n, the function h_n is integrable on $[0, 1]$, and the limit function is also integrable. One might therefore expect that

$$(*) \qquad \lim_{n\to\infty} \int_0^1 h_n = \int_0^1 \Big(\lim_{n\to\infty} h_n \Big).$$

However, the integral of h_1 over $[0, 1]$ is equal to 1, and because the entire "spike" of h_n occurs within the same interval, Exercise 7.13 shows that

$$\int_0^1 h_n = 1 \qquad \text{for all } n \geq 1,$$

so the left-hand side of $(*)$ is equal to 1. But the pointwise limit is the zero function, so the right-hand side of $(*)$ is equal to 0. Equation $(*)$ is false in this example! □

This example already suggests that pointwise convergence is too weak a notion of convergence. However, even worse things can happen: It is possible for the pointwise limit of a sequence of integrable functions not to be integrable at all.

Example 11.4 Let $(a_k)_{k=1}^{\infty}$ be a sequence that enumerates the rational numbers in the interval $[0, 1]$, see Theorem 3.19. For $n \in \mathbf{N}$, define

$$A_n = \bigcup_{k=1}^{n} \{a_k\};$$

the set A_n consists of "the the first n terms of the sequence (a_k). Finally, let $f_n : [0, 1] \to \mathbf{R}$ be the characteristic function of A_n. Viewing these graphs as a movie, we start with the zero function and at time n move one point on the graph—the point over a_n—up to height 1. Each function f_n is integrable, since f_n is identically zero except at finitely many points. However, the pointwise limit is the characteristic function of $\mathbf{Q} \cap [0, 1]$, which is not integrable as we saw in Chapter 7. □

Uniform Convergence

Pointwise convergence is inadequate because the fundamental operations of calculus—extraction of limits, integration, and differentiation—are *local* in character: They depend not on a function value at a point,

but on the behavior of a function on open intervals. If the sequence (f_n) converges pointwise on I, then for every $x \in I$ and every $\varepsilon > 0$, there exists an N such that

$$f_n(x) = f(x) + A(\varepsilon) \quad \text{for } n \geq N.$$

If a different point y is chosen, then for the same ε a larger N may be required, and if x ranges over even an arbitrarily small interval there may exist no upper bound on the corresponding N.

In hindsight, then, we cannot expect continuity (for example) to be inherited by the limit of a pointwise convergent sequence of functions. To remedy this situation, we introduce a stronger criterion of "convergence".

Definition 11.5 Let (f_n) be a sequence of functions on I that converges[1] pointwise to a function $f : I \to \mathbf{R}$, and let $X \subset I$. The sequence is said to *converge uniformly* to f on X if, for every $\varepsilon > 0$, there exists an index N such that

$$f_n = f + A(\varepsilon) \quad \text{on } X \text{ for } n \geq N. \tag{11.3}$$

If I is an interval and if $(f_n) \to f$ uniformly on every closed, bounded interval in I, then we say the convergence is *uniform on compacta*.

If $(f_n) \to f$ uniformly on I, then *a fortiori* the convergence is uniform on every non-empty subset of I. Uniform convergence has pointwise convergence built in, but is a stronger condition. Intuitively, "the same N works for all x," or "uniform convergence is to pointwise convergence as uniform continuity is to pointwise continuity." The latter is not a tautology; rather, it means the terminology is well-chosen. The *concepts*—not merely the names we have given them—are analogous.

Uniform convergence has a geometric interpretation. If $f : I \to \mathbf{R}$ is a function and $\varepsilon > 0$, then we define the *ε-tube* about the graph to be

$$\{(x, y) \in \mathbf{R}^2 : |f(x) - y| < \varepsilon,\ x \in I\},$$

namely the set of points that are within a vertical distance of ε from the graph of f.

To say (f_n) converges to f uniformly means that for every ε-tube about the graph of f, there is an N such that the graph of f_n lies within the tube for all $n \geq N$. Said yet another way, the sequence

[1] It is the *sequence* that converges, not the functions!

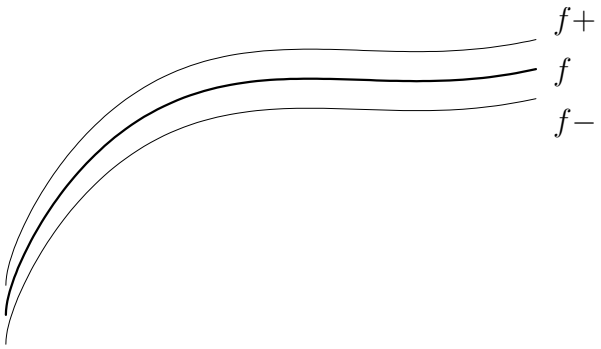


Figure 11.3: The ε-tube about the graph of a function.

(f_n) converges uniformly to f on I if the maximum vertical distance between the graphs of f_n and f can be made arbitrarily small. For later use we state this precisely:

Proposition 11.6. *Let (f_n) be a sequence that converges pointwise to f on I, and for each $n \in \mathbf{N}$ set*

$$a_n = \sup_{x \in I} |f(x) - f_n(x)|.$$

Then $(f_n) \to f$ uniformly on I iff $(a_n) \to 0$.

The proof is immediate from the definitions, and is left as an exercise. Proposition 11.6 is useful because the a_n may often be calculated easily.

Example 11.7 Let (f_n) be the sequence in Example 11.1. As we saw, this sequence converges pointwise to the characteristic function of $\{0\}$, that is

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } 0 < x \leq 1 \end{cases}$$

However, the convergence is not uniform even on the half-open interval $(0, 1]$ where f vanishes identically. In the notation of Proposition 11.6,

$$a_n = \sup\{f_n(x) \mid 0 < x \leq 1\} = 1 \qquad \text{for all } n,$$

so (a_n) does not converge to 0.

By contrast, consider an arbitrary closed interval of the form $[\delta, 1] \subset (0, 1]$. Since f_n vanishes identically on the interval $[\frac{1}{n}, 1]$, we find that

$a_n = 0$ as soon as $n > 1/\delta$. Thus, the sequence (f_n) converges uniformly to the zero function on each interval $[\delta, 1] \subset (0, 1]$. Since every closed subinterval of $(0, 1]$ is contained in an interval of the form $[\delta, 1]$, we see that $(f_n) \to 0$ uniformly on compacta in $(0, 1]$. Note carefully that removing a single point was not enough to "fix" the problem at 0, while removing an arbitrarily small interval *was* sufficient. □

Example 11.8 Let $\phi : \mathbf{R} \to \mathbf{R}$ be a *bounded* function, and define

$$f_n(x) = \frac{1}{n}\phi(nx), \qquad \text{for } x \in \mathbf{R}.$$

Geometrically, the graph of f_n is the graph of ϕ "shrunk" by a factor of n. Clearly $(f_n) \to 0$ uniformly on $\mathbf{R}$, since if $|\phi| \leq M$ on $\mathbf{R}$, then $|a_n| \leq M/n$ for each n.

Suppose in addition that ϕ is differentiable. Then f_n is differentiable for each n, and $f_n'(x) = \phi'(nx)$. Unless ϕ' is a constant function, the sequence (f_n') fails to converge even pointwise. In other words, uniform convergence of a sequence of differentiable functions does not even imply pointwise convergence of the sequence of derivatives. □

Despite the last example, some properties of the terms of a sequence (f_n), such as continuity and integrability, are inherited by a uniform limit.

Theorem 11.9. *Suppose (f_n) is a sequence of continuous functions on I that converge uniformly to f. Then f is continuous.*

In words, a uniform limit of continuous functions is continuous.

Proof. We wish to show that if $x \in I$, then $f - f(x) = o(1)$ at x. The trick is to write

$$f - f(x) = \big(f - f_N\big) + \big(f_N - f_N(x)\big) + \big(f_N(x) - f(x)\big). \tag{11.4}$$

Fix $\varepsilon > 0$, and use uniform convergence to choose N such that $f - f_N = A(\varepsilon/3)$ on I. Because f_N is continuous on I, there exists a neighborhood of x on which $f_N - f_N(x) = A(\varepsilon/3)$. On this neighborhood each of the three terms on the right in (11.4) is $A(\varepsilon/3)$, so their sum is $A(\varepsilon)$. Since $\varepsilon > 0$ was arbitrary, we have shown that $f - f(x) = o(1)$ at x. □

Theorem 11.10. *Let (f_n) be a sequence of integrable functions on $[a, b]$, that converges uniformly to f. Then the limit function f is integrable on $[a, b]$, and*

$$\lim_{n\to\infty} \int_a^b f_n = \int_a^b f = \int_a^b \Big(\lim_{n\to\infty} f_n\Big).$$

Proof. The strategy for showing the limit function is integrable is similar to the proof of Theorem 11.9: The limit function is everywhere close to an integrable function, so its upper and lower integrals cannot differ much. It is then straightforward to show the integral of the limit has the "correct" value.

Fix $\varepsilon > 0$, and choose N such that

$$|f(x) - f_N(x)| < \frac{\varepsilon}{b-a} \qquad \text{for all } x \in [a,b]. \tag{11.5}$$

By writing this inequality as

$$f_N(x) - \frac{\varepsilon}{b-a} \;<\; f(x) \;<\; f_N(x) + \frac{\varepsilon}{b-a} \qquad \text{for all } x \in [a,b],$$

and using the definitions of lower and upper sums, it is clear that

$$L(f_N, P) - \varepsilon \le L(f, P) \le U(f, P) \le U(f_N, P) + \varepsilon \tag{11.6}$$

for every partition P of $[a,b]$. Because f_N is integrable, there exists a partition P such that $U(f_N, P) - L(f_N, P) < \varepsilon$. For this partition, equation (11.6) implies $U(f,P) - L(f,P) < 3\varepsilon$, and since $\varepsilon > 0$ was arbitrary, f is integrable.

Now consider the sequence $(f - f_n)$, which converges uniformly to the zero function. Fix $\varepsilon > 0$ and choose N as in equation (11.5). Then

$$\left| \int_a^b f - \int_a^b f_n \right| = \left| \int_a^b (f - f_n) \right| \le \int_a^b |f - f_n| < \varepsilon$$

for $n \ge N$. This means $\lim_{n\to\infty} \int_a^b f_n = \int_a^b f = \int_a^b \left(\lim_{n\to\infty} f_n \right)$. □

There is, as already suggested, no analogous result for derivatives. The reason is that a function with small absolute value can have large derivative. Careful examination of the next theorem shows the hypotheses are qualitatively different from those of Theorems 11.9 and 11.10. The continuity hypothesis (i) below can be weakened, but the statement given here is adequate for our purposes.

Theorem 11.11. *Let (f_n) be a sequence of differentiable functions on an interval I, and assume in addition that*

(i) *Each function f_n' is continuous on I;*

(ii) *The sequence of* derivatives (f_n') *converges uniformly on compacta to a function* g*;*

(iii) *The original sequence* (f_n) *converges at a single point* $x_0 \in I$*.*

Then the original sequence (f_n) *converges uniformly on compacta to a differentiable function* f*, and* $f' = g = \lim f_n'$*.*

Proof. By the second fundamental theorem of calculus,

$$f_n(x) = f_n(x_0) + \int_{x_0}^{x} f_n'(t)\,dt \qquad \text{for all } x \in I.$$

Set $f(x_0) = \lim f_n(x_0)$. By (iii) and Theorem 11.10, the right-hand converges pointwise to

$$f(x) := f(x_0) + \int_{x_0}^{x} g(t)\,dt.$$

The function f is differentiable by the first fundamental theorem of calculus, and $f' = g$.

It remains to check that $(f_n) \to f$ uniformly on compacta. Fix $\varepsilon > 0$ and choose a closed, bounded interval $[a, b] \subset I$ that contains x_0. Note that if $x \in [a, b]$, then $|x - x_0| \leq (b - a)$. By (ii) and (iii), there exists $N \in \mathbf{N}$ such that $n > N$ implies

$$|f(x_0) - f_n(x_0)| < \frac{\varepsilon}{2} \text{ and } |g(x) - f_n'(x)| < \frac{\varepsilon}{2(b-a)} \quad \text{for all } x \in [a, b].$$

But this means that if $n > N$, then

$$\begin{aligned} |f(x) - f_n(x)| &\leq |f(x_0) - f_N(x_0)| + \int_{x_0}^{x} |g(t) - f_n'(t)|\,dt \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2(b-a)}|x - x_0| < \varepsilon, \end{aligned}$$

for *all* $x \in [a, b]$, which completes the proof. □

11.2 Series of Functions

Just as numerical infinite sums are limits of sequences of partial sums, infinite series of functions are limits of sequences of partial sums. Suppose $(f_k)_{k=0}^{\infty}$ is a sequence of functions on I, and define the sequence (s_n)

of *partial sums* by

$$s_n(x) = \sum_{k=0}^{n} f_k(x). \tag{11.7}$$

If the sequence (s_n) converges uniformly to f, then (f_k) is said to be *uniformly summable* on I, and we write

$$f = \sum_{k=0}^{\infty} f_k.$$

Power series, which have the form

$$\sum_{k=0}^{\infty} a_k (x - x_0)^k \quad \text{for some } x_0 \in \mathbf{R},$$

are the prototypical examples, but are by no means the only interesting ones. A power series arises from a sequence (f_k) in which f_k is a monomial of degree k (possibly zero) for each k. Not every polynomial series is a power series; an example is

$$\sum_{k=0}^{\infty} (x^k - x^{2k}),$$

though this is a *difference* of two power series. Generally, a sequence of polynomials can have surprising properties. We will see shortly that every continuous function $f : [0, 1] \to \mathbf{R}$ is the limit of a sequence of polynomials. This is remarkable because a general continuous function is differentiable nowhere, while a polynomial is infinitely differentiable everywhere.

Uniform Summability and Interchange of Limits

Theorem 11.10 says that if (f_k) is a uniformly summable sequence of integrable functions on some interval I, and if $[a, b] \subset I$, then $\sum_k f_k$ is integrable on $[a, b]$, and

$$\int_a^b \left(\sum_{k=0}^{\infty} f_k \right) = \sum_{k=0}^{\infty} \left(\int_a^b f_k \right).$$

Similarly, Theorem 11.11 says that if (f_k) is a sequence of continuously differentiable functions that is summable at a single point, and if the sequence of derivatives is uniformly summable, then

$$\left(\sum_{k=0}^{\infty} f_k\right)' = \sum_{k=0}^{\infty} f_k'.$$

When these equations hold, we say that the series $\sum_k f_k$ can be integrated or differentiated *term-by-term*. For finite sums there is no issue, since integration and differentiation are linear operators, but an infinite sum involves a second limit operation. These equations assert that under appropriate hypotheses, two limit operations can be interchanged. Remember that we are primarily interested in power series. The equations above guarantee that on an interval where a power series is uniformly summable, it may be treated computationally as if it were a polynomial.

Before we can use these results to full advantage, we need a simple criterion for determining when a sequence of functions is uniformly summable, and in particular when a power series is uniformly summable. The desired simple, general criterion for uniform summability is known as the *Weierstrass M-test*, compare Proposition 11.6.

Theorem 11.12. *Let (f_k) be a sequence of functions on I, and set*

$$a_k = \sup_{x \in I} |f_k(x)| \geq 0.$$

If (a_k) is a summable sequence of real numbers, then (f_k) is uniformly summable on I.

Proof. Fix $x \in I$. The sequence $\big(f_k(x)\big)$ is absolutely summable since $|f_k(x)| \leq a_k$ and (a_k) is summable. Let $f : I \to \mathbf{R}$ be the pointwise sum of the sequence (f_k). To see that the partial sums converge uniformly to f, write

$$\left| f(x) - \sum_{k=0}^{n} f_k(x) \right| = \left| \sum_{k=n+1}^{\infty} f_k(x) \right| \leq \sum_{k=n+1}^{\infty} |f_k(x)| \leq \sum_{k=n+1}^{\infty} a_k.$$

This is the tail of a convergent sum, which converges to 0 as $n \to \infty$ independently of x. □

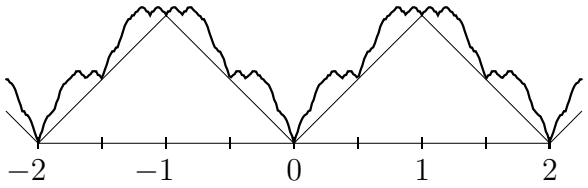


Figure 11.4: A Weierstrass nowhere-differentiable function.

Before we consider power series in detail, here is the long-promised example of a continuous function that is nowhere differentiable. This function was discovered by Weierstrass in the mid-19th Century, and shocked mathematicians of the day, who were accustomed to regarding "functions" as being differentiable more-or-less everywhere.

Example 11.13 Let $\mathrm{cb} : \mathbf{R} \to \mathbf{R}$ be the Charlie Brown function, and consider the sequence $(f_k)_{k=1}^{\infty}$ defined by

$$f_k(x) = 4^{-k}\,\mathrm{cb}(4^k x).$$

Geometrically, the graph of f_k is the graph of cb "zoomed out" by a factor of 4^k; in particular, f_k is periodic with period 4^{-k}, and the maximum a_k of f_k is 4^{-k}. The sequence $(4^{-k})_{k=0}^{\infty}$ is geometric, with ratio $1/4$, hence is summable. By the M-test, the sequence (f_k) is uniformly summable on $\mathbf{R}$, and since each function f_k is continuous, the function

$$F = \sum_{k=1}^{\infty} f_k$$

is also continuous. Because cb is even and has period 2, F is even and has period 2, so it suffices to prove F is nowhere-differentiable in $[0, 1]$. The recursive structure of F is the key.

There is an easy geometric reason for non-differentiability: Subtracting off the "first mode" gives $F(x) - \mathrm{cb}(x) = \frac{1}{4}F(4x)$, see Figure 11.4. In words, up to addition of a piecewise linear term, the graph of the Weierstrass function is self-similar. In Chapter 8, we saw that zooming in on the graph of a differentiable function gives the tangent line. However, self-similarity of the graph guarantees that no matter how much we zoom in, the graph does not become more nearly linear. As it stands, this argument does not provide a proof, because we have not considered what happens at "corners" of the summands, though it

is probably quite believable that F fails to be differentiable at points where a partial sum has a corner.

An analytic proof of non-differentiability is not difficult if we organize our knowledge carefully. It is enough to show that the Newton quotients fail to exist:

$$\text{For every } x \in [0,1],\ \lim_{h\to 0} \frac{F(x+h)-F(x)}{h} \quad \text{does not exist.}$$

Fix a positive integer n and consider the nth partial sum of the series, and the nth tail:

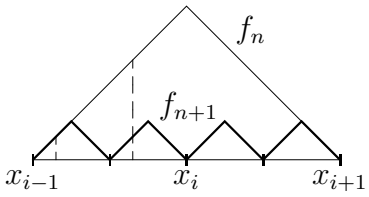
$$F_n(x) = \sum_{k=0}^{n} 4^{-k}\,\mathrm{cb}(4^k x), \qquad t_n(x) = \sum_{k=n+1}^{\infty} 4^{-k}\,\mathrm{cb}(4^k x).$$

The function f_{n+1} is periodic with period $4^{-(n+1)}$, so t_n is as well. Thus

$$F(x \pm 4^{-n}) - F(x) = F_n(x \pm 4^{-n}) - F_n(x) \qquad \text{for all } n. \tag{11.8}$$

Now fix x; we will construct a sequence $(h_n)_{n=0}^{\infty}$ with $h_n = \pm 4^{-n}$ such that the corresponding Newton quotients have no limit as $n \to \infty$.

Consider the points $x_i = i2^{-(2n+1)}$ for $0 \le i \le 2^{2n+1}$. The nth and $(n+1)$st summands look like this:



The period of f_n is 4^{-n}, and the distance between adjacent points of the partition is one-half this distance, which is twice the period of f_{n+1}. The point x lies in at least one subinterval of the form $[x_{i-1}, x_i]$, and the length of this interval is $2 \cdot 4^{-(n+1)}$. Consequently, there is at least one choice of sign so that $x \pm 4^{-n}$ and x lie in $[x_{i-1}, x_i]$; this defines the sequence $(h_n)_{n=0}^{\infty}$. The positions of a typical such pair are depicted as dashed lines.

For each summand f_k with $k \le n$, the points x and $x + h_n$ lie in an interval on which f_k is linear, so the Newton quotient of f_k between x and $x + h_n$ is ± 1. By (11.8),

$$\frac{F(x+h_n)-F(x)}{4^{-n}} = \frac{F_n(x+h_n)-F_n(x)}{4^{-n}} = \sum_{k=0}^{n} \frac{f_k(x+h_n)-f_k(x)}{4^{-n}}$$

is an integer for each n. Moreover, increasing n by 1 does not change the value of any of the summands on the right, but does add another term, which changes the quotient by ± 1.

This shows that each Newton quotient is an integer, but that consecutive quotients differ. Consequently, the sequence of Newton quotients has no limit, i.e., F is not differentiable at x, for all $x \in [0,1]$. □

We chose a particular scaling (each summand 1/4 the size of the previous) in order to simplify the proof of non-differentiability. It is possible to use similar ideas with sequences that scale differently, obtaining more examples.

11.3 Power Series

Recall that a *formal power series* is an expression of the form

$$\sum_{k=0}^{\infty} a_k (x-a)^k;$$

the *coefficients* constitute a sequence (a_k), and the point a is the *center* of the series. Associated to a formal power series is a sequence of approximating polynomials, which are obtained by truncating the series:

$$p_n(x) = \sum_{k=0}^{n} a_k (x-a)^k.$$

A formal power series defines a function f whose domain is the set of x for which the series converges, and this always contains at least the point a, at which the value is a_0. It is a *convention* that $(x-a)^0 = 1$ even when $x = a$; this does not mean that $0^0 = 1$.

Theorem 11.14. *Suppose the power series $\sum_{k=0}^{\infty} a_k (x-a)^k$ converges at x_0, and set $|x_0 - a| = \eta$. If $|x-a| < \eta$, then the power series converges at x, and the convergence is uniform on compacta in $(a-\eta, a+\eta)$.*

Proof. It suffices to assume $a = 0$ (and hence define $\eta = |x_0|$); this amounts to making a translational change of variable. Suppose the power series $\sum_k a_k x^k$ converges at a point x_0. In particular, the sequence $(a_k x_0^k)_{k=0}^{\infty}$ converges to 0, which in turn implies the sequence is bounded: There exists a real number M such that

$$|a_k x_0^k| \leq M \quad \text{for all } k \in \mathbf{N}.$$

Now suppose $|x| < |x_0|$, that is, x is closer to the center of the series than x_0. Then $0 < \rho := |x|/|x_0| < 1$, and

$$\begin{aligned} |f(x) - p_n(x)| &= \left| \sum_{k=n+1}^{\infty} a_k x^k \right| = \left| \sum_{k=n+1}^{\infty} a_k x_0^k \cdot \frac{x^k}{x_0^k} \right| \\ &\leq \sum_{k=n+1}^{\infty} |a_k x_0^k| \cdot \rho^k \leq \sum_{k=n+1}^{\infty} M\rho^k = \frac{M}{1-\rho}\,\rho^{n+1} \end{aligned}$$

for every $n \in \mathbf{N}$. This upper bound can be made arbitrarily small by taking n sufficiently large; in other words, the power series converges at x.

It remains to show the convergence is uniform on compacta in $(a - \eta, a + \eta)$. Fix a positive number $\delta < \eta$, and set $\rho = \delta/\eta < 1$. If $|x - a| \leq \delta$, then the argument above shows that

$$|f(x) - p_n(x)| \leq \frac{M}{1-\rho}\,\rho^{n+1}$$

independently of x, and this can be made arbitrarily small because $\rho < 1$. □

Consequently, if a power series converges at a single point $x_0 \neq a$, then it converges on an open interval $I = (a - \eta, a + \eta)$, and converges uniformly on every closed subinterval of I. On I, the series represents a differentiable function whose derivative can be computed term-by-term. Contrapositively, if a power series *diverges* at x_0, then it also diverges for all x with $|x - a| > |x_0 - a|$.

Given a power series, the set of real numbers is partitioned into two sets; those at which the series converges, and those at which it diverges. Theorem 11.14 implies the former is an interval centered at a; it is, naturally, called the *interval of convergence* of the power series. Let $R \geq 0$ be the supremum of $|x - a|$ over the interval of convergence. The interval of convergence of a power series must be one of the following:

- The singleton $\{a\}$ (if $R = 0$);
- A bounded interval centered at a (if $0 < R < \infty$);
- All of $\mathbf{R}$ (if $R = \infty$).

Example 11.16 demonstrates that the interval of convergence may be open, half-open, or closed.

The radius of convergence of a power series depends only on the coefficient sequence (a_k) and not on the center. There is a formula, due to Hadamard, that gives the radius in terms of the coefficients:

$$\frac{1}{R} = \lim_{n\to\infty} \sup_{k\geq n} \sqrt[k]{|a_k|}.$$

We do not need this generality, and will instead develop simpler formulas that work only for certain sequences of coefficients, but including all those we shall encounter.

Theorem 11.15. *Let (a_n) be a sequence, and suppose*

$$L = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right|$$

exists. If $L > 0$, then $R = 1/L$ is the radius of convergence of the power series

$$\sum_{n=0}^{\infty} a_n (x-a)^n.$$

If $L = 0$, the power series converges for all $x \in \mathbf{R}$.

Proof. We use the hypothesis to compare the power series with a geometric series. Suppose $0 < |x-a| < R$, with R as in the theorem. Then

$$\lim_{n\to\infty} \left| \frac{a_{n+1}(x-a)^{n+1}}{a_n(x-a)^n} \right| = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| \cdot |x-a| < 1.$$

Because the limit is strictly smaller than 1, it may be written $1-2\varepsilon$, with $\varepsilon > 0$. From the definition of convergence, there exists $N \in \mathbf{N}$ such that

$$\left| \frac{a_{n+1}(x-a)^{n+1}}{a_n(x-a)^n} \right| < 1-\varepsilon =: \rho \qquad \text{for } n > N.$$

Induction on m proves that

$$|a_{N+m}(x-a)^{N+m}| \leq |a_N(x-a)^N|\,\rho^m, \qquad m \geq 0.$$

Aside from the first N terms, the power series is therefore dominated by the terms of a convergent geometric series.

An entirely analogous argument shows that if $|x-a| > R$, then the power series diverges; thus R is the radius. □

Theorem 11.15 is called the *ratio test*. When applicable, it is often the simplest way to calculate the radius of convergence of a power series. Note that the ratio does not give any information about convergence at $a \pm R$. Convergence at endpoints must *always* be checked separately, by hand.

Example 11.16 The examples here are centered at 0 to simplify notation. The ratio test applies to each series, but calculation of the radius is left to you.

- The power series $\sum_{k=0}^{\infty} x^k$ has radius 1. When $x = \pm 1$ (the endpoints of the interval of convergence), the series diverges, as its terms do not have limit zero. Consequently, the interval of convergence is $(-1, 1)$.
- The series $\sum_{k=0}^{\infty} \frac{x^k}{k}$ has radius 1. At $x = -1$, the series converges by the alternating series test, while at $x = 1$ the series is harmonic, and therefore divergent. The interval of convergence is the half-open interval $[-1, 1)$.
- The series $\sum_{k=0}^{\infty} \frac{x^k}{k^2}$ has radius 1. At each endpoint, the series converges absolutely by comparison with the 2-series. The interval of convergence is the closed interval $[-1, 1]$.
- The series $\sum_{k=0}^{\infty} \frac{x^k}{k!}$, see equation (11.1), has radius $+\infty$, so the interval of convergence is $\mathbf{R}$.
- The series $\sum_{k=0}^{\infty} k^k x^k$ has radius 0, so the "interval" of convergence is the singleton $\{0\}$.

As these examples demonstrate, to find the interval of convergence, you first calculate the radius of convergence, then check the endpoints. The only time you need not check the endpoints is when there *are* no endpoints, either because $R = 0$ or $R = +\infty$. □

Real-Analytic Functions

Theorem 11.14 says that if a power series centered at a has radius $R > 0$, then on the interval $(a - R, a + R)$, the series represents a continuous function of x. Indeed, if $x_0 \in (a - R, a + R)$, then there is a $\delta > 0$ such that

$$[x_0 - \delta, x_0 + \delta] \subset (a - R, a + R),$$

and the sequence of partial sums converges uniformly on $[x_0 - \delta, x_0 + \delta]$, hence represents a continuous function on this interval, and can be integrated term-by-term. Since x_0 was arbitrary, the power series represents a continuous function on $(a-R, a+R)$ and can be integrated termwise on this interval.

In fact, much more is true; the power series obtained by termwise differentiation has the same radius as the original series, which can be used to show the original power series represents a *differentiable* function on $(a-R, a+R)$. This innocuous fact can be bootstrapped, showing that a power series represents an *infinitely differentiable* function on its open interval of convergence.

We will not prove these claims in full generality, though it is easy to modify the arguments below, replacing occurrences of the ratio test with Hadamard's formula for the radius of a power series. (Naturally, one must also prove that Hadamard's formula generally gives the radius of a power series.)

Theorem 11.17. *Let* $\sum_{k=0}^{\infty} a_k(x-a)^k$ *be a power series such that*

$$\frac{1}{R} = \lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right|$$

exists and is positive, and let $f : (a-R, a+R) \to \mathbf{R}$ *be the sum of the series. Then* f *is differentiable; in fact, the termwise derived series*

$$\sum_{k=1}^{\infty} k a_k (x-a)^{k-1} = \sum_{k=0}^{\infty} (k+1) a_{k+1} (x-a)^k \tag{11.9}$$

has radius R*, and represents* f' *on* $(a-R, a+R)$*.*

Proof. By the ratio test, the derived series has radius

$$\lim_{n\to\infty} \left| \frac{(n+1)a_{n+1}}{(n+2)a_{n+2}} \right| = \lim_{n\to\infty} \left(\frac{n+1}{n+2} \right) \left(\left| \frac{a_{n+1}}{a_{n+2}} \right| \right) = \lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right|,$$

so on the interval $(a-R, a+R)$, the series (11.9) represents a continuous

function g that can be integrated termwise. For each $x \in (a-R, a+R)$,

$$\begin{aligned}\int_a^x g(t)\,dt &= \int_a^x \left(\sum_{k=1}^{\infty} k a_k (t-a)^{k-1}\right) dt \\ &= \sum_{k=1}^{\infty}\left(\int_a^x k a_k (t-a)^{k-1}\,dt\right) = \sum_{k=1}^{\infty} a_k (t-a)^k \Big|_{t=a}^{t=x} \\ &= \sum_{k=1}^{\infty} a_k (x-a)^k = f(x) - f(a).\end{aligned}$$

Since g is continuous, the second fundamental theorem implies $f' = g$ on $(a-R, a+R)$. □

The logic is a little delicate, and bears one more repetition. We begin with a power series $\sum_k a_k x^k$ whose radius (as given by the ratio test) is $R > 0$. The termwise derived series $\sum_k k a_k x^{k-1}$ has the same radius of convergence, and by a separate argument represents the derivative of the original series. This establishes our principal goal of the chapter, to prove that convergent power series can be manipulated as if they were polynomials with infinitely many terms. But if this were not enough, we can bootstrap the argument: If f is represented by a power series on an open interval I, then f' is also represented by a power series on the same interval. Consequently, f'' is represented by a power series on I, and so forth. Formally, induction shows that the function f is infinitely differentiable, and the successive derivatives may be found by repeated *termwise* differentiation.

Definition 11.18 Let $I \subset \mathbf{R}$ be an open interval. A function $f : I \to \mathbf{R}$ is said to be *real analytic* if, for each $a \in I$, there is a $\delta > 0$ and a power series centered at a that converges to $f(x)$ for all $|x-a| < \delta$.

Many functions of elementary mathematics are real analytic. The reason for the apparently convoluted definition is the fact of life that a single power series is not generally sufficient to represent an analytic function on its entire domain. Real analytic functions will provide us with some striking evaluations of infinite sums, and will be crucial to our construction of the trig functions in Chapter 13. For now we are content to establish the basic arithmetic properties of real analytic functions.

Theorem 11.19. *Let f and g be real analytic functions on an interval I. Then the functions $f+g$ and fg are real analytic on I, and f/g is analytic on each interval where g is non-vanishing.*

Proof. Real analyticity is a local property, so the content of the theorem is really that a sum, product, or quotient of convergent power series can be represented by a convergent power series (the latter provided the denominator is non-zero). For simplicity, assume all series are centered at 0; substituting $x - a$ for x takes care of the general case. Write

$$f(x) = \sum_{i=0}^{\infty} a_i x^i, \qquad g(x) = \sum_{j=0}^{\infty} b_j x^j,$$

and let $p_n(x)$ and $q_n(x)$ denote the respective nth partial sums. Because each series converges for some non-zero x, there is an $\eta > 0$ such that each series is uniformly summable on the interval $[-\eta, \eta]$.

Sums are done in Exercise 11.3. Products are handled with the Cauchy formula for the product of absolutely summable series, see Chapter 4. Formally, we multiply the series and collect terms of degree k:

$$\left(\sum_{i=0}^{\infty} a_i x^i\right)\left(\sum_{j=0}^{\infty} b_j x^j\right) = \sum_{k=0}^{\infty}\left(\sum_{i=0}^{k} a_i b_{k-i}\right) x^k. \tag{11.10}$$

Because a power series converges absolutely inside its interval of convergence, the series on the right is equal to the product of the series on the left on $[-\eta, \eta]$.

Let g be analytic, and assume $g(0) \neq 0$. In order to prove that $1/g$ is analytic at 0, we write g as a power series,

$$g(x) = a_0 + a_1 x + a_2 x^2 + \cdots = \sum_{k=0}^{\infty} a_k x^k,$$

and seek a series

$$h(x) = \sum_{n=0}^{\infty} b_k x^k$$

such that $g(x)h(x) = 1$ for all x in some neighborhood of 0. Writing out the coefficients of the product and equating with the power series of 1, we have

$$b_0 = \frac{1}{a_0}, \qquad \sum_{i=0}^{k} a_i b_{k-i} = 0 \quad \text{for all } k \geq 1.$$

This system of infinitely many equations can be solved recursively. First rewrite it as

$$b_0 = \frac{1}{a_0}, \qquad b_k = -\frac{1}{a_0}\sum_{i=1}^{k} a_i b_{k-i} \quad \text{for } k \geq 1.$$

Then $k = 1$ gives $b_1 = -(a_1/a_0)b_0 = -a_1/(a_0)^2$, $k = 2$ gives

$$b_2 = -\frac{1}{a_0}(a_1 b_1 + a_2 b_0) = -\frac{1}{{a_0}^3}(a_0 a_2 - a_1^2),$$

and so on. To check convergence of the reciprocal series, use the "geometric series trick": Write $g(x) = a_0\big(1 - \phi(x)\big)$, with $\phi(x) = O(x)$. Then

$$\frac{1}{g(x)} = \frac{1}{a_0} \cdot \frac{1}{1 - \phi(x)} = \frac{1}{a_0}\sum_{k=0}^{\infty} \phi(x)^k,$$

in some neighborhood of 0. See Example 11.23 for a concrete application. □

Power series and O notation interact very nicely, which explains the use of O notation in many symbolic computer packages. We state the result only for power series centered at 0; there is an obvious modification for series centered at a.

Corollary 11.20. *If $f(x) = \sum_{k=0}^{\infty} a_k x^k$ is analytic at 0, then for each positive integer n we have*

$$f(x) = \sum_{k=0}^{n} a_k x^k + O(x^{n+1}).$$

Proof. By Theorem 11.19,

$$f(x) - \sum_{k=0}^{n} a_k x^k = \sum_{k=n+1}^{\infty} a_k x^k = x^{n+1}\sum_{j=0}^{\infty} a_{n+j+1} x^j,$$

and the sum on the right is real-analytic, hence $O(1)$ near 0. □

The following result is called the *identity theorem* for power series. The practical consequence is that two power series centered at a define the same function if *and only if* their coefficients are the same. Again, we state the result only for series centered at 0.

Corollary 11.21. *If* $\sum_{k=0}^{\infty} a_k x^k \equiv 0$ *near* 0, *then* $a_k = 0$ *for all* k.

Proof. We prove the contrapositive. Assume not all the a_k are zero, and let a_n be the first non-zero coefficient. By Theorem 11.19 and Corollary 11.20,

$$\sum_{k=0}^{\infty} a_k x^k = x^n \sum_{j=0}^{\infty} a_{n+j} x^j = x^n \big(a_n + O(x)\big).$$

Since $a_n \neq 0$, the term in parentheses is non-vanishing on some neighborhood of 0, so the series is non-vanishing on some deleted interval about 0. □

Theorem 11.19 is the basis of calculational techniques for finding power series of reciprocals. Two examples will serve to illustrate useful methods.

Example 11.22 Let $g(x) = 1 - x$, which is clearly analytic and non-zero at 0. The coefficients of g are $a_0 = 1$, $a_1 = -1$. The reciprocal has power series

$$h(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \cdots,$$

where $b_0 = 1$ and $0 = a_0 b_k + a_1 b_{k-1} = b_k - b_{k-1}$ for all $k \geq 1$. We conclude immediately that all the coefficients of $1/g$ are equal to 1:

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots = \sum_{k=0}^{\infty} x^k.$$

This is nothing but the geometric series formula. □

The geometric series trick in the proof of Theorem 11.19 is a nice calculational means of finding the coefficients of a reciprocal series if only the first few coefficients are needed.

Example 11.23 Suppose $f(x) = 1 - \frac{1}{2}x^2 + \frac{1}{4!}x^4 + O(x^6)$ (a certain interesting series starts this way) and that we wish to find the power series of $1/f$. Write $f(x) = 1 - \phi(x)$, then use the geometric series formula:

$$\frac{1}{1-\phi(x)} = 1 + \phi(x) + \phi(x)^2 + \phi(x)^3 + \cdots$$

This procedure is justified because $\phi(0) = 0$, so we have $|\phi(x)| < 1$ on some neighborhood of 0 by continuity. Algebra gives

$$\begin{aligned}
\phi(x) &= \frac{1}{2}x^2 - \frac{1}{4!}x^4 + O(x^6) \\
\phi(x)^2 &= \frac{1}{4}x^4 + O(x^6) \\
\phi(x)^3 &= O(x^6) \\
\frac{1}{f(x)} &= 1 + \frac{1}{2}x^2 + \frac{5}{24}x^4 + O(x^6).
\end{aligned}$$

We cannot get any more terms with the information given, but if f is known up to $O(x^n)$ then we are guaranteed to find the reciprocal up to the same order. □

11.4 Approximating Sequences

Power series are not the only interesting approximating sequences. In this section we introduce a couple of more specialized examples.

Picard Iterates

Recursively defined sequences of functions arise naturally in approximating solutions of differential equations. The general first-order differential equation in one space variable may be written

$$y'(t) = f\big(t, y(t)\big), \qquad y(t_0) = y_0. \tag{11.11}$$

The second equation is called an *initial value*, and is often regarded as specifying the value of y at time t_0. Integrating both sides of this equation from t_0 to t gives the equivalent integral equation

$$y(t) = y(t_0) + \int_{t_0}^{t} f\big(s, y(s)\big)\, ds. \tag{11.12}$$

The right-hand side of this equation may be regarded as a function of t that also *depends on the function* y.[2] In other words, there is an

[2]The right-hand side depends upon f, but we regard f as fixed in this discussion.

operator P that maps functions to functions; a function y is mapped to the function Py defined by

$$Py(t) = y(t_0) + \int_{t_0}^{t} f\big(s, y(s)\big)\, ds.$$

A solution of (11.12) is exactly a function y such that $Py = y$, namely a fixed point of P. Motivated by recursively defined numerical sequences, we hope to find fixed points of P by starting with an initial guess and iterating P. Our initial guess is the constant function y_0, and we set $y_{n+1} = Py_n$ for $n \geq 0$, namely

$$y_{n+1}(t) = y_0 + \int_{t_0}^{t} f\big(s, y_n(s)\big)\, ds. \tag{11.13}$$

The terms of the sequence $(y_n)_{n=0}^{\infty}$ are called *Picard iterates* for the initial-value problem (11.11). Under a fairly mild restriction on f, the sequence of Picard iterates converges uniformly on some neighborhood of t_0 to a fixed point of P.

Example 11.24 To get a feel for Picard iterates in a specific example, consider the initial-value problem

$$y' = y, \qquad y(0) = 1, \tag{11.14}$$

whose solution is the natural exponential function. Here $f(t, y) = y$, so $f\big(s, y_n(s)\big) = y_n(s)$. We make the initial guess $y_0(t) = 1$ for all t in some interval about 0. Equation (11.13) gives

$$\begin{aligned}
y_1(t) &= 1 + \int_0^t y_0(s)\, ds = 1 + t \\
y_2(t) &= 1 + \int_0^t y_1(s)\, ds = 1 + t + \frac{t^2}{2 \cdot 1} \\
y_3(t) &= 1 + \int_0^t y_2(s)\, ds = 1 + t + \frac{t^2}{2 \cdot 1} + \frac{t^3}{3 \cdot 2 \cdot 1},
\end{aligned}$$

and so forth. It seems we are recovering our old friend from (11.1), and indeed it is easy to check by induction on n that

$$y_n(t) = \sum_{k=0}^{n} \frac{t^k}{k!} \qquad \text{for } n \geq 0.$$

Formally, if we iterate an infinite number of times we obtain the power series (11.1). Differentiating this series as if it were a polynomial, we find that the derivative of each summand in (11.1) is the preceding summand, so (at least formally) $y' = y$. □

Approximate Identities

In this section we study the "convolution product" of functions, an operation of considerable importance in signal processing. The convolution product also has important theoretical applications to approximation. We will prove the striking fact that a continuous function on a closed, bounded interval can be uniformly approximated by a sequence of polynomials.

Formally, the *convolution product* of f and g is defined by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)\,dt. \tag{11.15}$$

However, this improper integral is not generally convergent. Our first task is to restrict attention to a suitable vector space of functions.

A function $f \in \mathscr{F}(\mathbf{R}, \mathbf{R})$ is *compactly supported* if there exists an R in $\mathbf{R}$ such that $f(x) = 0$ for $|x| > R$. In words, f is identically zero outside some closed interval. The set of continuous, compactly supported functions on $\mathbf{R}$ is denoted $\mathscr{C}_c^0(\mathbf{R})$. This set is a vector subspace of $\mathscr{F}(\mathbf{R}, \mathbf{R})$: If f and g are continuous and compactly supported, then $f + g$ and cf are clearly continuous and compactly supported.

Lemma 11.25. *If f and g are in $\mathscr{C}_c^0(\mathbf{R})$, then so is $f * g$.*

Proof. Suppose $f(x) = 0$ for $|x| > R_1$ and $g(x) = 0$ for $|x| > R_2$. We claim that if $|x| > R_1 + R_2$, then $f(t)g(x-t) = 0$ for all t, so surely $(f * g)(x) = 0$. But the reverse triangle inequality says that if $|t| \le R_1$, then

$$|x - t| \ge \big||x| - |t|\big| > (R_1 + R_2) - R_1 = R_2.$$

In other words, if $f(t) \neq 0$, then $g(x-t) = 0$. □

A continuous, compactly supported function is a model of a "signal" that is only non-zero for a bounded interval of time, or a "response curve of a filter". Convolving a signal f with a filter response g gives the output signal.

Example 11.26 Let $g = (b-a)^{-1}\chi_{[a,b]}$ be a "unit impulse" in $[a, b]$. Then $g(x-t) = 1/(b-a)$ if $x - b \le t \le x - a$ and is zero otherwise, so

$$(f * g)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)\,dt = \frac{1}{b-a}\int_{x-b}^{x-a} f(t)\,dt$$

for all $f \in \mathscr{C}_c^0(\mathbf{R})$. The integral on the right is the average value of f on $[x-b, x-a]$.

More generally, if $g(x) > 0$ for $x \in (a, b)$ and is zero elsewhere, and if g encloses one unit of area, then $f * g$ may be viewed as the result of "averaging" f over intervals of length $b-a$. □

Despite its seemingly strange definition, the convolution product satisfies two beautiful identities:

Proposition 11.27. *Let f_1, f_2, and f_3 be continuous and compactly supported. Then $f_1 * f_2 = f_2 * f_1$ and $(f_1 * f_2) * f_3 = f_1 * (f_2 * f_3)$.*

In words, convolution is commutative and associative. The proof of commutativity is left to you, see Exercise 11.14. Associativity requires a result from integration in two variables ("Fubini's theorem"), and is mentioned only for conceptual reasons; we do not use associativity in this book.

The Dirac δ-Function

Dirac's δ-*function* is a fictitious "function" with the following properties:

- $\displaystyle\int_{-\infty}^{\infty} \delta(t)\, dt = 1.$
- $\delta(t) = 0$ if $t \neq 0$.

For example, $\delta(x-t)$ is a "unit impulse concentrated at x", so formally

$$(f * \delta)(x) = \int_{-\infty}^{\infty} f(t)\delta(x-t)\, dt = f(x);$$

convolving with δ is the identity mapping.

Unfortunately, the properties above are logically incompatible: If a function satisfies the second property, then its integral is 0. However, physicists and electrical engineers have found this "function" to be extremely useful in their work, and if pressed by a mathematician will usually reconcile the two properties above by saying, "Yes, but $\delta(0) = \infty$". Engineers are even willing to regard the δ-function as the derivative of the "Heaviside step function", defined by $H(x) = 0$ if $x < 0$, $H(x) = 1$ if $x \geq 0$!

The utility of the δ-function strongly suggests that a precise mathematical concept is lurking. Physicists started using the δ-function in the early days of quantum mechanics, and within 20 years mathematicians had found at least three rigorous interpretations. We introduce one of

these, the "approximate identity". Rather than think of the Dirac δ as a single function, we construct a sequence that approximately satisfies the conditions above.

Definition 11.28 A sequence of non-negative functions (δ_n) is an *approximate identity* if:

- For all $n \in \mathbf{N}$, δ_n is integrable, and $\displaystyle\int_{-\infty}^{\infty} \delta_n = 1$.
- For every $\beta > 0$, $\displaystyle\lim_{n\to\infty} \int_{-\beta}^{\beta} \delta_n = 1$.

The second condition formalizes the idea that "the integrals concentrate at 0". We have assumed $\delta_n \geq 0$ only for simplicity; with more work, the proofs below can be extended to remove the non-negativity hypothesis. The formal calculation that $f * \delta = f$ takes the following rigorous form; again, you should think of the Goethe quote!

Theorem 11.29. *Let $f \in \mathscr{C}_c^0(\mathbf{R})$, and let (δ_n) be an approximate identity. The sequence (f_n) defined by $f_n = f * \delta_n$ converges uniformly to f.*

Proof. By commutativity of the convolution product, we have

$$f_n(x) = \int_{-\infty}^{\infty} f(x-t)\delta_n(t)\,dt, \qquad f(x) = \int_{-\infty}^{\infty} f(x)\delta_n(t)\,dt$$

for all n. Linearity of the integral, non-negativity of the δ_n, and the triangle inequality imply

$$\begin{aligned} \big|f(x) - f_n(x)\big| &= \left| \int_{-\infty}^{\infty} \big[f(x) - f(x-t)\big]\,\delta_n(t)\,dt \right| \\ &\leq \int_{-\infty}^{\infty} \Big|f(x) - f(x-t)\Big|\,\delta_n(t)\,dt. \end{aligned}$$

We wish to show that this can be made small independently of x by choosing n sufficiently large. For every β, the last integral may be split into

$$\int_{-\beta}^{\beta} \Big|f(x) - f(x-t)\Big|\,\delta_n(t)\,dt + \int_{|t|\geq\beta} \Big|f(x) - f(x-t)\Big|\,\delta_n(t)\,dt.$$

The idea is that for small t, the increment of f is small (so the first term is small), while for large t the contribution is small because the δ_n concentrate at 0.

Fix $\varepsilon > 0$. Because f is continuous and compactly supported, f is uniformly continuous on $\mathbf{R}$: There exists $\beta > 0$ such that $|t| \le \beta$ implies $|f(x) - f(x-t)| < \varepsilon$. It also follows that f is bounded: There exists $M > 0$ such that $|f(x) - f(x-t)| \le M$ for *all* $x, t \in \mathbf{R}$. Use the second property in Definition 11.28 to choose N such that

$$\int_{|t|\ge\beta} \delta_n(t)\,dt < \frac{\varepsilon}{M} \qquad \text{for } n \ge N.$$

With these choices, if $n \ge N$, then

$$\int_{-\beta}^{\beta} \underbrace{\big|f(x) - f(x-t)\big|}_{<\varepsilon}\,\delta_n(t)\,dt + \int_{|t|\ge\beta} \underbrace{\big|f(x) - f(x-t)\big|}_{\le M}\,\delta_n(t)\,dt$$
$$< \int_{-\beta}^{\beta} \varepsilon\delta_n(t)\,dt + \int_{|t|\ge\beta} M\delta_n(t)\,dt \le \varepsilon + M\cdot\frac{\varepsilon}{M} = 2\varepsilon$$

independently of x. □

The Weierstrass Approximation Theorem

We have seen two ways in which a sequence of polynomials can be used to approximate a function f. The first, Lagrange interpolation (Chapter 3), constructs polynomials that agree with f at specified points. However, there is no guarantee that the approximation is good over the entire domain of f. The second, partial sums of a power series, is only applicable to real-analytic functions. Chapter 14 is devoted to careful study of analytic functions and their approximating polynomials. In this section we mention a third type of approximation, due to K. Weierstrass, in which a sequence of polynomials is used to approximate a general continuous function *uniformly*. The sequence is constructed by convolving with a suitably chosen approximate identity.

Theorem 11.30. *Let $f : [a,b] \to \mathbf{R}$ be continuous. There exists a sequence (p_n) of polynomials such that $(p_n) \to f$ uniformly on $[a,b]$.*

Proof. By adding a linear polynomial (as in the proof of the mean value theorem) we may assume that $f(a) = f(b) = 0$: We can approximate f by polynomials iff we can approximate

$$g(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b-a}(x-a)$$

by polynomials. Second, we may substitute $x = a + (b-a)u$, reducing to the case $[a,b] = [0,1]$. In more detail, if we write $\widetilde{\phi}(u) = \phi(x)$, then $(p_n) \to f$ uniformly on $[a,b]$ iff $(\widetilde{p_n}) \to \widetilde{f}$ uniformly on $[0,1]$. Thus it suffices to prove the theorem for a continuous function $f : [0,1] \to \mathbf{R}$ that satisfies $f(0) = f(1) = 0$.

We first construct an approximate identity consisting of piecewise polynomial functions. Let c_n be defined by

$$c_n \int_{-1}^{1} (1-x^2)^n \, dx = 1,$$

and set

$$\delta_n(x) = \begin{cases} c_n(1-x^2)^n & \text{if } -1 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

In Chapter 15 we will evaluate c_n explicitly, but for now the following estimate is enough:

$$\begin{aligned} \frac{1}{c_n} &= \int_{-1}^{1} (1-x^2)^n \, dx = 2\int_0^1 (1-x^2)^n \, dx \\ &> 2\int_0^{1/\sqrt{n}} (1-x^2)^n \, dx > 2\int_0^{1/\sqrt{n}} (1-nx^2)\, dx = \frac{4}{3\sqrt{n}}; \end{aligned}$$

The second inequality is a consequence of Exercise 2.5, part (b). We deduce immediately that (δ_n) is an approximate identity, since

$$0 \le \delta_n(x) \le \tfrac{3\sqrt{n}}{4}(1-x^2)^n,$$

and the upper bound converges uniformly to 0 off $[-\beta,\beta]$ for all $\beta > 0$.

Now set $p_n = f * \delta_n$; because $f = 0$ outside $[0,1]$, we have

$$p_n(x) = \int_{-\infty}^{\infty} f(t)\delta_n(x-t)\, dt = \int_0^1 f(t)\delta_n(x-t)\, dt.$$

If $x \in [0,1]$, then $x - t \in [-1,1]$ for $0 \le t \le 1$, so $\delta_n(x-t) = c_n\big(1-(x-t)^2\big)^n$. The integrand $f(t)\delta_n(x-t)$ is therefore a polynomial in x whose coefficients are continuous functions of t; upon integrating in t from 0 to 1, we find that p_n is a polynomial in x.

By Theorem 11.29, $(p_n) \to f$ uniformly on $[0,1]$. □

Several points are worth emphasizing. First, the approximate identity is explicit, so for a specific function f there is an effective computational way of finding the approximating polynomials. Second, the

space of continuous functions is large and complicated, while the space of polynomials is simple and explicit. Conceptually, the Weierstrass theorem is analogous to the density of **Q** (a simple set of numbers) in **R** (an enormous, complicated set). Finally, a general continuous function is differentiable *nowhere*, yet is uniformly approximated on every closed, bounded interval by *smooth* functions.

Exercises

Exercise 11.1 Let $\phi : \mathbf{R} \to \mathbf{R}$ be as in Example 11.3, and let $(h_n)_{n=0}^{\infty}$ be the sequence defined by $h_n(x) = n\phi(nx)$. Carefully sketch the graphs of ϕ and h_n ($n = 1, 2, 3$) on a single set of axes. Prove that (h_n) converges pointwise to the zero function. ◇

Exercise 11.2 Let $\phi : \mathbf{R} \to \mathbf{R}$ be continuous, and define (f_n) by

$$f_n(x) = \frac{1}{n}\phi(x), \qquad \text{for } x \in \mathbf{R}.$$

(a) Prove that $(f_n) \to 0$ uniformly on compacta.

(b) Prove that $(f_n) \to 0$ uniformly if and only if f is bounded.

(c) Suppose

$$\sum_{k=0}^{\infty} a_k x^k$$

converges pointwise on **R** (i.e., has infinite radius of convergence). Prove that the partial sums converge uniformly on **R** if and only if the sum is finite.

◇

Exercise 11.3 Prove that if f and g are real analytic in a neighborhood of 0, then $f + g$ is real analytic. Hint: All the necessary estimates can be found in earlier parts of the text. ◇

Exercise 11.4 Compute the power series of $1/(1+x)$ and $1/(1+x^2)$. What are the radii of convergence? Use the sum and product formulas for series to verify that

$$\frac{1}{1-x} + \frac{1}{1+x} = \frac{2}{1-x^2} \quad \text{and} \quad \frac{1}{1-x} \cdot \frac{1}{1+x} = \frac{1}{1-x^2}.$$

◇

Exercise 11.5 Use the reciprocal trick to compute the series of

$$\frac{1}{1-x^2+x^4}$$

up to and including terms of degree 6. ◇

Exercise 11.6 Use the product and geometric series formulas to compute the power series of

$$\frac{1+x}{1-x}.$$

What is the radius? ◇

Exercise 11.7 Let

$$f(x) = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Compute the radius of convergence, and prove that $f' = f$ on its interval of convergence. Find the power series of $\big(f(x)-1\big)/x$, and use your answer to compute the reciprocal series, $x/\big(f(x)-1\big)$, up to (and including) terms of degree 3.
Hint: When computing powers in the reciprocal trick, you needn't carry terms whose degree is larger than 3. ◇

Exercise 11.8 Let $f(x)$ be the series of the preceding problem. Use the product formula for series and the binomial theorem to show that

$$f(x)f(y) = f(x+y) \quad \text{for all } x,\, y \in \mathbf{R}.$$

Does this confirm anything you already know? ◇

Exercise 11.9 The geometric series formula says

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots = \sum_{k=0}^{\infty} x^k.$$

What does this equation say (formally) if $x = 1$? What if $x = -1$? $x = 2$? Do any of these formulas make sense? Explain. ◇

Exercise 11.10 Differentiate the geometric series formula—as given in the previous exercise—on the open interval $(-1, 1)$. Use the result to find closed-form expressions for the power series

$$f(x) = \sum_{k=1}^{\infty} kx^k, \quad g(x) = \sum_{k=1}^{\infty} k^2 x^k, \qquad -1 < x < 1.$$

Similarly, integrate the geometric series from 0 to x, and express the result in closed form. For which x is the resulting formula true? ◇

Exercise 11.11 Find all real x such that the following is correct:

$$\frac{1}{x}+\frac{1}{x^2}+\frac{1}{x^3}+\cdots=\sum_{k=1}^{\infty}\left(\frac{1}{x}\right)^k=\frac{\frac{1}{x}}{1-\frac{1}{x}}=\frac{1}{x-1}.$$

Can we conclude that $-\sum_{k=0}^{\infty}x^k=-\frac{1}{1-x}=\sum_{k=1}^{\infty}\left(\frac{1}{x}\right)^k$? ◇

Exercise 11.12 Find all real x such that $\sum_{k=0}^{\infty}e^{kx}$ converges, and express the sum in closed form. ◇

Exercise 11.13 Find the radii of the following power series:

(a) $\displaystyle\sum_{k=1}^{\infty}\frac{(-x)^k k!}{k^k}$ (b) $\displaystyle\sum_{k=1}^{\infty}x^{k!}$

◇

Exercise 11.14 Prove the commutativity part of Proposition 11.27. ◇

Exercise 11.15 Let $(f_n)_{n=1}^{\infty}$ be a sequence of non-decreasing functions, and assume the series $\sum_n |f_n(0)|$ and $\sum_n |f_n(1)|$ converge.

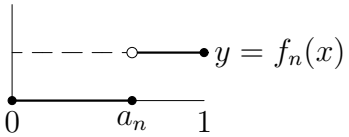
(a) Prove that the series $\displaystyle\sum_{n=1}^{\infty}f_n(x)$ is convergent for each $x\in[0,1]$.

(b) Prove that the function $f:=\sum_n f_n$ is non-decreasing.

(c) Prove that the convergence in part (a) is uniform on $[0,1]$.

(d) Let $(a_n)_{n=1}^{\infty}$ be a sequence in $[0,1]$, and assume the terms are distinct: $a_n\neq a_m$ if $n\neq m$. Define

$$f_n(x)=\begin{cases}0 & \text{if } 0\le x\le a_n\\ 2^{-n} & \text{if } a_n<x\le 1\end{cases}$$



Prove that f is discontinuous at $x\in[0,1]$ iff $x=a_k$ for some k.

(e) Show that there exists an *increasing* function $f:[0,1]\to[0,1]$ that is discontinuous at x for every $x\in\mathbf{Q}\cap[0,1]$.

◇

Chapter 12

Log and Exp

Aside from "pathological" or piecewise defined examples, the natural logarithm function log and its inverse, the natural exponential function exp, are the first non-algebraic functions studied in this book. Their importance cannot be summarized in just a few sentences, though their ubiquity throughout the natural sciences can be explained fairly simply: The natural exponential function arises in any situation where the rate of change of some quantity is proportional to the quantity itself. Populations grow at a rate roughly proportional to their size; money accrues interest at a rate proportional to the principle; to a good approximation, many chemical reactions proceed at a rate proportional to the concentrations of the reactants; radioactive nuclei decay at random, so the number of decays per second is proportional to the number of nuclei in a sample. In each situation, the amount of "stuff" present at time t will be well-approximated by an exponential function of t.

Logarithms are a convenient language in any situation where a quantity varies over a large range, speaking in terms of ratios. The energy carried by a sound wave or the concentration of hydrogen ions in a solution are quantities that in realistic situations range over many orders of magnitude. Logarithmic units (such as decibels or pH) make such quantities more manageable. (The loudest sound a human ear can tolerate without pain carries billions of times more energy than a whisper, for example, but we speak conveniently of 130 decibels versus 30 decibels.)

Logarithmic and exponential functions have axiomatic definitions (see below), that were historically the basis for their discovery. Our treatment inverts the historical order for logical reasons.

12.1 The Natural Logarithm

Historically, a "logarithm" is a function L, not identically zero, that converts multiplication into addition in the sense that

$$L(xy) = L(x) + L(y) \qquad \text{for all positive, real } x \text{ and } y. \tag{12.1}$$

In school courses, it is granted that such functions exist, and this equation is taken as an axiom. We are now in a position to introduce and study functions with this property, but for us the previous equation will be a *theorem*.

The *natural logarithm* is the function $\log : (0, \infty) \to \mathbf{R}$ defined by

$$\log x = \int_1^x \frac{1}{t}\, dt \quad \text{for } x > 0. \tag{12.2}$$

Properties of the integral immediately imply that $\log(1) = 0$. The second fundamental theorem shows that log is differentiable, and that

$$\log' x = \frac{1}{x}, \qquad x > 0. \tag{12.3}$$

Consequently log is increasing on $(0, \infty)$, and in particular is positive iff $x > 1$, see Figures 12.1 and 12.2.

Theorem 12.1. $\log(xy) = \log x + \log y$ *for all* $x,\ y > 0$.

Proof. If x and y are positive real numbers, then

$$\begin{aligned} \log(xy) &= \int_1^{xy} \frac{dt}{t} = \int_1^x \frac{dt}{t} + \int_x^{xy} \frac{dt}{t} \\ &= \int_1^x \frac{dt}{t} + \int_1^y \frac{du}{u} = \log(x) + \log(y). \end{aligned}$$

The change of limits on the second integral is justified by setting $t = xu$ (remember that "x is a constant") and invoking Exercise 7.13. □

In Leibniz notation, the logarithm property (12.1) follows from scale invariance of the integrand dt/t:

$$\frac{dt}{t} = \frac{d(xu)}{(xu)} = \frac{du}{u} \qquad \text{for all } x > 0.$$

It is not difficult to show (Exercise 12.1) that up to an overall multiplied constant, dt/t is the only *continuous* integrand with the requisite invariance property.

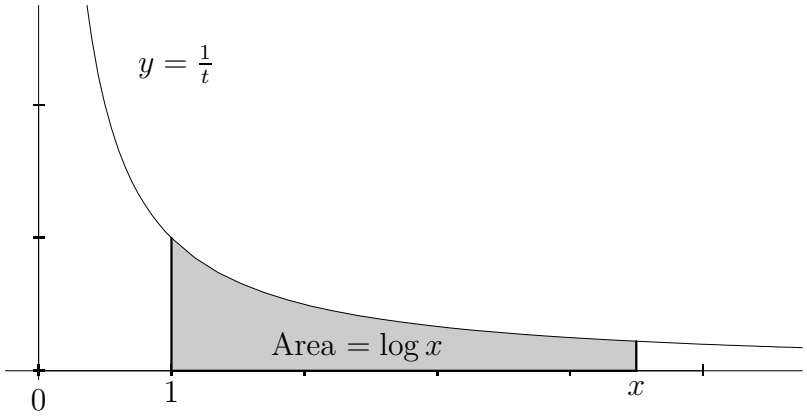


Figure 12.1: The definition of the natural logarithm as an integral.

Equation (12.1) implies $\log(1/x) = -\log x$ for all $x > 0$, and (by induction on p) that $\log(x^p) = p \log x$ for all $p \in \mathbf{N}$. Setting $x = y^{1/q}$ for $q \in \mathbf{N}$ and assembling the previous observations shows that

$$\log(y^r) = r \log y \qquad \text{for all real } y > 0, \text{ all } r = \tfrac{p}{q} \text{ rational.} \tag{12.4}$$

The real number $\log 2$ is positive, and $\log(2^n) = n \log 2$ for all integers n. By the Archimedean property of $\mathbf{R}$, the log function attains arbitrarily large (positive and negative) values. Since log is continuous, it has the intermediate value property, so we conclude that log maps $(0, \infty)$ *onto* $\mathbf{R}$. Because log is increasing, every real number is the logarithm of a *unique* positive real number.

Note carefully that the tangent lines to the graph of log become arbitrarily close to horizontal (because $\log' x = 1/x$), but the graph has no horizontal asymptote!

12.2 The Natural Exponential

Historically, an *exponential function* is a function $E : \mathbf{R} \to \mathbf{R}$ such that

$$E(x + y) = E(x)E(y) \qquad \text{for all } x, y \in \mathbf{R}, \tag{12.5}$$

cf. (12.1). Recall that we defined the *natural exponential function* $\exp : \mathbf{R} \to (0, \infty)$ to be the inverse of log, the natural logarithm. The reason for the definition is historically rooted:

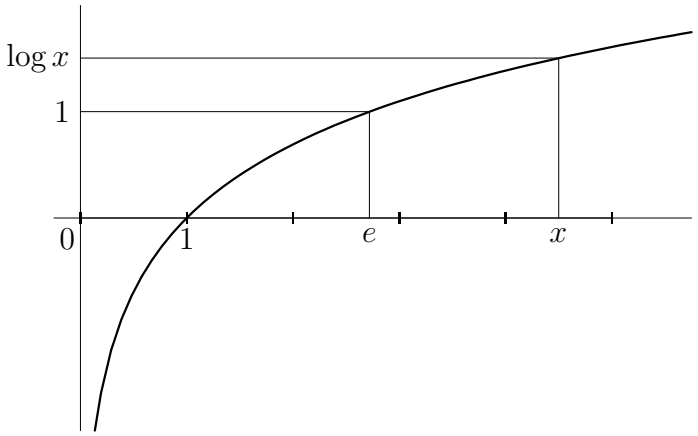


Figure 12.2: The graph of the natural logarithm function.

Lemma 12.2. *If $L : (0,\infty) \to \mathbf{R}$ is a logarithm function, then its inverse $E : \mathbf{R} \to (0,\infty)$ is an exponential function.*

Proof. A logarithm, by definition, satisfies the identity $L(xy) = L(x) + L(y)$ for all positive x and y. If we write $x = E(u)$, $y = E(v)$, and apply E to the logarithm identity, we find that

$$E(u) \cdot E(v) = xy = E\big(L(x) + L(y)\big) = E(u+v) \qquad \text{for all } u \text{ and } v,$$

which is the characteristic property of an exponential function. □

The Number e

The number $e := \exp 1$ is one of the most prominent constants in mathematics. Note that by definition, $\log e = 1$; e is the unique real number for which the region in Figure 12.1 has unit area. Exercise 7.17 (d) shows that $2 < e < 4$; the actual value is roughly

$$e = 2.718281828459045\ldots,$$

see Corollary 12.8.

12.3 Properties of exp and log

Equation (12.4) says $y^r = \exp(r \log y)$ for all real $y > 0$, all rational r. The expression on the left has a purely algebraic definition (involving

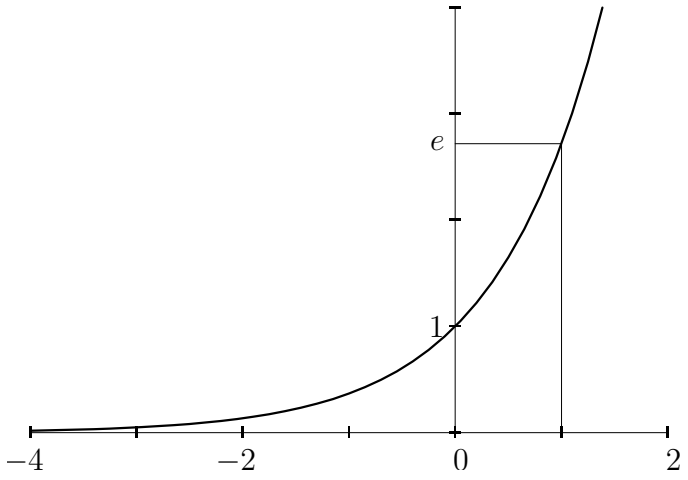


Figure 12.3: The graph of the natural exponential function.

nothing but multiplication of real numbers), while the right-hand side is not purely algebraic, but has the advantage of being defined for all *real* r. We are led to *define*

$$x^r = \exp(r \log x) \qquad \text{for } x > 0,\ r \in \mathbf{R}. \tag{12.6}$$

In particular,

$$e^x = \exp x \quad \text{for all } x \in \mathbf{R}.$$

If $b > 0$, the exponential function to the base b, denoted $\exp_b$, is defined by

$$\exp_b x = \exp(x \log b) \quad \text{for all } x \in \mathbf{R}.$$

The name will be justified shortly.

An exponential function is proportional to its own derivative, and exponential functions are characterized by this property:

Theorem 12.3. *Let $b > 0$ be fixed. The function $\exp_b$ is differentiable, and*

$$\exp_b' x = (\log b) \exp_b x \quad \textit{for all } x \in \mathbf{R}.$$

Conversely, if $k \in \mathbf{R}$ and if f is a differentiable function satisfying the differential equation $f' = kf$, then $f(x) = f(0)e^{kx}$ for all $x \in \mathbf{R}$.

Proof. Recall that $\exp' = \exp$; the proof is instructive, and is repeated here. First, $\log(\exp x) = x$ for all $x \in \mathbf{R}$. Second, log is differentiable and has non-vanishing derivative, so its inverse function is

differentiable. It is therefore permissible to differentiate the equation $x = \log(\exp x)$ with respect to x. The chain rule gives

$$1 = \log'(\exp x)\, \exp' x = \frac{1}{\exp x}\, \exp' x \qquad \text{for all } x,$$

proving $\exp' = \exp$. If $k \in \mathbf{R}$, the chain rule implies $\frac{d}{dx} e^{kx} = ke^{kx}$. Since $\exp_b x = \exp(x \log b)$ by definition, the derivative formula is immediate.

Conversely, suppose $f : \mathbf{R} \to \mathbf{R}$ is a differentiable function such that $f' = kf$. Define g by $g(x) = f(x)/e^{kx}$; this is sensible because exp is non-vanishing on $\mathbf{R}$. The function g is differentiable as a quotient of differentiable functions, and the quotient rule implies

$$g'(x) = \frac{e^{kx} f'(x) - ke^{kx} f(x)}{(e^{kx})^2} = \frac{f'(x) - kf(x)}{e^{kx}} = 0 \quad \text{for all } x$$

since $f'(x) = kf(x)$ for all x by hypothesis. But this means g is a constant function, and setting $x = 0$ shows $g(x) = f(0)$ for all x. □

Corollary 12.4. *Let $r \in \mathbf{R}$ be fixed. If $f(x) = x^r$ for $x > 0$, then f is differentiable, and $f'(x) = rx^{r-1}$. In Leibniz notation,*

$$\frac{d}{dx} x^r = rx^{r-1} \qquad \textit{for all } r \in \mathbf{R}.$$

The proof is left to you, Exercise 12.6. Theorem 12.3 also implies some familiar algebraic properties of exp:

Theorem 12.5. *For all x, $y \in \mathbf{R}$, $e^{x+y} = e^x e^y$ and $e^{xy} = (e^y)^x$.*

Proof. Fix $y \in \mathbf{R}$, and consider the function $f : \mathbf{R} \to \mathbf{R}$ defined by $f(x) = e^{x+y}$. By the chain rule, f is differentiable, and $f' = f$. Since $f(0) = e^y$, Theorem 12.3 implies $e^{x+y} = e^x e^y$ for all x.

To prove the second assertion, fix y and define $g : \mathbf{R} \to \mathbf{R}$ by $g(x) = (e^y)^x$. Applying the first part of Theorem 12.3 with $b = e^y$, we see that

$$g'(x) = \log(e^y)(e^y)^x = yg(x).$$

The second part of Theorem 12.3 implies $g(x) = e^{xy}$, since $g(0) = 1$. As a fringe benefit, we have shown that $(e^y)^x = (e^x)^y$, since each term is equal to e^{xy}. □

Theorem 12.5 justifies the name "exponential function to the base b":

$$\exp_b(x) = \exp(x \log b) = \big(\exp(\log b)\big)^x = b^x \qquad \text{for all } x \in \mathbf{R}.$$

Easy modifications of the proof establish the identities

$$b^{x+y} = b^x b^y, \qquad b^{xy} = (b^x)^y \qquad \text{for all } b > 0,\ x,\ y \in \mathbf{R}.$$

If you attempt to prove these identities directly from equation (12.6), you will be impressed by the simplicity of the argument just given. One of the powers of calculus is its ability to encode algebraic information in differential equations that have unique solutions.

You may have wondered more than once why we do not define $0^0 = 1$. Here is a limit of the form 0^0 that is not equal to 1:

$$\lim_{x \to 0} \exp(\alpha/x^2)^{x^2} = \lim_{x \to 0} \big(e^{(\alpha/x^2)}\big)^{x^2} = e^\alpha \qquad \text{for all } \alpha < 0,$$

by Theorem 12.5. It is left to you to find functions f and g such that $\lim(f, 0) = \lim(g, 0) = \lim(f^g, 0) = 0$.

The inverse of $\exp_b$ is the *logarithm to the base* b, and is denoted $\log_b$:

$$y = \log_b x \quad \text{iff} \quad x = b^y = \exp_b y.$$

Theorem 12.5 implies $\log_b(x^y) = y \log_b x$ for all $x > 0$, $y \in \mathbf{R}$.

The next result says the logarithm to the base b is proportional to the natural logarithm. For this reason, logarithm functions other than the natural logarithm appear only rarely in mathematics.

Proposition 12.6. *If $b > 0$, then $\log_b x = (\log x)/(\log b)$ for all $x \in \mathbf{R}$. The function $\log_b$ is differentiable, and*

$$\log_b' x = \frac{1}{x \log b} \qquad \textit{for all } x > 0.$$

Proof. To say $y = \log_b x$ means $x = b^y = \exp(y \log b)$. Taking the natural logarithm of this equation gives $y = (\log x)/(\log b)$. The statement about derivatives follows immediately. □

Two Representations of exp

The characterization of exp as the solution of the differential equation $f' = f$ satisfying $f(0) = 1$ implies a couple of striking, non-trivial representations.

Theorem 12.7. $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ *for all* $x \in \mathbf{R}$.

Proof. The power series on the right has $a_k = 1/k!$, so the ratio test implies that the radius of convergence is

$$\lim_{k\to\infty} \frac{a_k}{a_{k+1}} = \lim_{k\to\infty} \frac{(k+1)!}{k!} = \lim_{k\to\infty} k+1 = \infty.$$

The associated function f is defined for all $x \in \mathbf{R}$, and is differentiable. The derivative f' is found by differentiating term-by-term:

$$f'(x) = \sum_{k=0}^{\infty} \frac{1}{k!}\, kx^{k-1} = \sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!} = f(x)$$

for all x. Since $f(0) = 1$ (all terms but the first vanish), Theorem 12.3 implies $f(x) = e^x$ for all x. □

Corollary 12.8. $e = \sum_{k=0}^{\infty} \frac{1}{k!} = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots$

In order to turn the result of this corollary into an effective computational fact, we need to know the size of the error if we approximate e by adding up finitely many terms of this series. Merely summing the first four terms shows that $2.6\bar{6} < e$, which is already a substantial improvement over $2 < e$. A good numerical estimate is found in Exercise 12.23.

Theorem 12.9. $e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$ *for all* $x \in \mathbf{R}$.

Proof. Remember that x is fixed as the limit is taken. We begin by explicitly allowing "n" to take arbitrary positive *real* values, rather than integer values. The "change of variable" $h = 1/n$ converts the

desired limit to

$$\begin{aligned}\lim_{h\to 0^+}(1+xh)^{1/h} &= \lim_{h\to 0^+}\exp\left[\frac{1}{h}\log(1+xh)\right] \\ &= \lim_{h\to 0^+}\exp\left[x\frac{\log(1+xh)-\log 1}{xh}\right] \quad \text{since } \log 1 = 0 \\ &= \exp\left[x\lim_{h\to 0^+}\frac{\log(1+xh)-\log 1}{xh}\right] \quad \text{continuity of exp.}\end{aligned}$$

Setting $\eta = xh$ and noticing that the limit term is the Newton quotient for the natural logarithm shows the previous expression is equal to

$$\exp\left[x\lim_{\eta\to 0}\frac{\log(1+\eta)-\log 1}{\eta}\right] = \exp\left[x\ \log'(1)\right] = \exp x,$$

and this is e^x by definition. □

Theorem 12.9 characterizes the natural exponential function as a limit of geometric growth. If, for example, x is the annual interest rate on a savings account, and there are n compoundings per year, then the multiplier on the right gives the factor by which the savings increase over one year. As the number of compoundings per year grows without bound, the balance does not become infinite in a finite time. Instead, if \$1 is allowed to accrue interest with continuous compounding, then in the time it would take the savings to double without compounding, the balance increases to \$2.72 (rounded to the nearest penny).

Exercises

Exercise 12.1 Let $f : (0,\infty) \to \mathbf{R}$ be a continuous function, and define

$$L(x) = \int_1^x f(t)\,dt.$$

Prove that if L satisfies (12.1), then there exists a real k such that $f(t) = k/t$ for all $t > 0$. ◇

Exercise 12.2 Let a, b, and c be positive real numbers. Is it more sensible to agree that a^{b^c} is equal to $(a^b)^c$ or to $a^{(b^c)}$, or does it matter? ◇

Exercise 12.3 Prove that $x^{\log y} = y^{\log x}$ for all $x, y > 0$. ◇

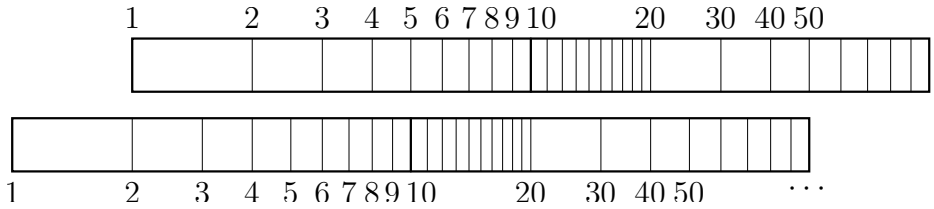


Figure 12.4: A ruler marked with common (base 10) logarithms.

Exercise 12.4 Explain how you would use two logarithmic scales, as in Figure 12.4, to multiply numbers. ◇

Exercise 12.5 Let u be a differentiable function. Find the derivative of $\exp \circ u$, and the derivative of $\log \circ |u|$ at points where u is non-zero. Write your formulas in both Newton and Leibniz notation. ◇

Exercise 12.6 Prove Corollary 12.4.
Hint: Begin with the definition of x^r for real r. ◇

Exercise 12.7 Let $f(x) = e^{-x}(x^2 - 1)$ for $x \in \mathbf{R}$.

(a) Sketch the graph of f, using information about the first two derivatives to determine intervals on which f is monotone, convex, and concave.
Suggestion: Introduce symbolic constants to simplify calculations.

(b) How many solutions does the equation $f(x) = a$ have? (Your answer will depend on a.)

◇

Exercise 12.8 If $f : (0, \infty) \to \mathbf{R}$ is defined by $f(x) = x^x$, find f'. Prove that f has a unique minimum, and find the location and value of the minimum. Hint: Write $f(x) = e^{u(x)}$ for an appropriate function u. ◇

Exercise 12.9 Evaluate the following limits:

(a) $\lim_{x \to 0^+} x \log x$ (b) $\lim_{x \to 0^+} x^x$ (c) $\lim_{x \to +\infty} x^{1/x}$ (d) $\lim_{x \to +\infty} (1 + x)^{1/x}$

◇

Exercise 12.10 Evaluate the following limits:

(a) $\lim_{x \to +\infty} \frac{x}{\log x}(x^{1/x} - 1)$ (b) $\lim_{x \to 0^+} \frac{e - (1 + x)^{1/x}}{x}$ ◇

Exercise 12.11 Let n be a positive integer. Evaluate

(a) $\lim_{x\to+\infty} \dfrac{(\log x)^n}{x}$ (b) $\lim_{x\to+\infty} \dfrac{x^n}{e^x} = \lim_{x\to+\infty} x^n e^{-x}$

Use your results to prove that for all $\alpha > 0$, $\log x = o(x^\alpha)$ near $+\infty$. Find a similar little-o expression for power and exponential functions. ◇

Exercise 12.12 Find the maximum value of $f_n(x) = x^n e^{-x}$ for $x \geq 0$. (In particular, you must prove that a maximum value exists.) ◇

Exercise 12.13 Prove that $\displaystyle\int_0^\infty e^{-t^2}\,dt$ converges. ◇

Exercise 12.14 Let $F_n(x) = \displaystyle\int_0^x t^n e^{-t}\,dt$.

(a) Use integration by parts (Exercise 10.12) to find a recursion formula for F_n in terms of F_{n-1}.

(b) Use part (a) and induction on n to prove that

$$F_n(x) = n!\Big(1 - e^{-x}\sum_{k=0}^{n}\frac{x^k}{k!}\Big).$$

(c) Evaluate the improper integral $\displaystyle\int_0^\infty t^n e^{-t}\,dt$

This integral is defined for all real $n > -1$, and is denoted $\Gamma(n+1)$. ◇

Exercise 12.15

(a) Use part (b) of the previous exercise and a change of variable to find a formula for

$$\int_0^x t^n e^{-\alpha t}\,dt, \qquad \alpha \in \mathbf{R}.$$

(b) Prove that the improper integral $\displaystyle\int_0^1 (\log u)^n\,du$ converges.

(c) Use the change of variable $u = e^t$ to evaluate the improper integral of part (b).

◇

Exercise 12.16 Let $f : (0, +\infty) \to \mathbf{R}$ be an increasing function. Recall that f is integrable on $[a, b]$ for all $0 < a < b$.

(a) Prove that

$$\sum_{k=1}^{n-1} f(k) \leq \int_1^n f(t)\,dt \leq \sum_{k=2}^{n} f(k)$$

for all $n \in \mathbf{N}$. (A sketch should help. Compare Proposition 7.23.)

(b) Taking $f = \log$ in part (a), prove that

$$(n-1)! \leq e\Big(\frac{n}{e}\Big)^n \leq n!$$

for all $n \in \mathbf{N}$.

(c) Evaluate $\lim_{n\to\infty} \dfrac{(n!)^{1/n}}{n}$.

There is a much more precise estimate of $n!$ called *Stirling's formula*. ◇

Exercise 12.17 Determine (with proof, of course) which of the following converge; do not attempt to evaluate the sums!

(a) $\displaystyle\sum_{n=1}^{\infty} \frac{\log n}{n^{3/2}}$ (b) $\displaystyle\sum_{n=1}^{\infty} (-1)^n \frac{\log n}{n}$ (c) $\displaystyle\sum_{n=2}^{\infty} \frac{1}{n(\log n)^2}$

(d) $\displaystyle\sum_{n=1}^{\infty} \frac{(n+1)^n}{n^{n+1}}$ (e) $\displaystyle\sum_{n=1}^{\infty} \frac{n!}{n^n}$

Hint for (a): "Borrow" a small power of n to nullify the log. ◇

Exercise 12.18 Let n be a positive integer. ◇

Exercise 12.19 Define $f : \mathbf{R} \to \mathbf{R}$ by

$$f(x) = \begin{cases} \exp(-1/x^2) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Prove that $f^{(k)}(0)$ exists and is equal to zero for all $k \in \mathbf{N}$.
Suggestion: First use induction on the degree to prove that for every polynomial p,

$$\lim_{x\to 0} p(1/x)\, f(x) = 0.$$

Then show inductively that every derivative of f is of this form. Finally, show that no derivative of f can be discontinuous at 0. ◇

Exercise 12.20 Fix $b > 0$, and define a sequence $(x_n)_{n=0}^{\infty}$ by

$$x_0 = 1, \quad x_{n+1} = b^{x_n} \text{ for } n \geq 0.$$

Thus $x_1 = b$, $x_2 = b^b$, $x_3 = b^{b^b}$, and so forth. Prove that if $(x_n) \to \ell$, then $b^\ell = \ell$. Use this observation to determine the set of b for which the sequence converges. Then show that ℓ is an increasing function of b, and find the largest possible value of ℓ.

Two people are arguing. One says that if $b = \sqrt{2}$, then $\ell = 2$, since $\sqrt{2}^2 = 2$; the other says $\ell = 4$ because $\sqrt{2}^4 = 4$. Who—if either—is correct, and why? ◇

Exercise 12.21 Let n and x be positive integers. Prove that

$$n \leq \log_{10} x < n + 1 \quad \text{iff} \quad 10^n \leq x < 10^{n+1},$$

iff x is an integer having $n + 1$ digits. In words, the integer part of the base 10 logarithm of x is one less than the number of digits of x.

Which is larger, $2^{2^{2^{2^2}}}$ or $10^{10^{100}}$? How many digits does each number have? ◇

Exercise 12.22 The mother of all l'Hôpital's rule problems: Prove that

$$\lim_{x\to\infty} e^{e^{e^{x+e^{-(a+x+e^x+e^{e^x})}}}} - e^{e^{e^x}} = e^{-a}$$

for all $a \in \mathbf{R}$. ◇

Exercise 12.23 The error in using the first $n + 1$ terms of the series in Corollary 12.8 to estimate e is

$$(*) \qquad e - \sum_{k=0}^{n} \frac{1}{k!} = \sum_{k=n+1}^{\infty} \frac{1}{k!}.$$

(a) Show that $(n + m)! \geq (n + 1)^m n!$ for $m \geq 1$. (If you write out what this means for $m = 2$ the inductive proof should be clear.)

(b) Show that the error in $(*)$ is no larger than $1/(n \cdot n!)$.
Suggestion: Use part (a) and a geometric series.

(c) Use part (b) to show that $2.716\bar{6} < e < 2.7183\bar{3}$. You may use only the arithmetic operations on a calculator.

(d) How many terms suffice to give 20 decimals of accuracy? Give as small an answer as you can. (Surely 10^{20} terms suffice!)

◇

Exercise 12.24 Prove that e is irrational. Hint: Assume $e = p/q$ is rational, in lowest terms. By Exercise 12.23,

$$0 < \frac{p}{q} - \sum_{k=0}^{n} \frac{1}{k!} < \frac{1}{n(n!)} \quad \text{for each } n \in \mathbf{N}.$$

Take $n = q$ and deduce there is a positive integer smaller than $1/q$. (Remember: If $k \leq q$, then $k!$ divides $q!$ evenly.) ◇

Chapter 13

The Trigonometric Functions

Trigonometric functions are usually introduced via geometry, either as ratios of sides in a right triangle, or in terms of points on the unit circle. The approach taken here may, by contrast, seem opaque, even artificial. However, our aim is to define everything in terms of axioms of $\mathbf{R}$, so we shall give an analytic definition of the trig functions. In order to make contact with geometry, we must show that our definitions coincide with the familiar geometric definitions. These arguments will necessarily be geometric, but as their purpose is pedagogical (rather than logical) the reliance on geometry will not detract from the logical structure of the chapter.

13.1 Sine and Cosine

The exponential function exp is characterized by a first-order differential equation, namely it is the unique differentiable function $f : \mathbf{R} \to \mathbf{R}$ such that

$$f' = f \quad \text{and} \quad f(0) = 1. \tag{13.1}$$

The *uniqueness* assertion—equation (13.1) has *at most one* solution—used little more than the mean value theorem, while the *existence* part—(13.1) has *at least one* solution—required some additional machinery, either integration or power series. Our approach to the elementary circular trig functions is similar. The following definition relies implicitly on theorems that the given criteria do indeed uniquely define functions; these theorems will presently be formally stated and

proved. Uniqueness will be an easy argument using the mean value theorem, while existence will depend on power series.

Definition 13.1 The *sine* function $\sin : \mathbf{R} \to \mathbf{R}$ is the solution of the initial-value problem

$$f'' + f = 0, \qquad f(0) = 0, \quad f'(0) = 1. \tag{13.2}$$

The *cosine* function $\cos : \mathbf{R} \to \mathbf{R}$ is the solution of the initial-value problem

$$f'' + f = 0, \qquad f(0) = 1, \quad f'(0) = 0. \tag{13.3}$$

The tangent, cotangent, secant, and cosecant functions are defined (with their natural domains) to be ratios of sin and cos in the usual manner:

$$\tan = \frac{\sin}{\cos}, \quad \cot = \frac{\cos}{\sin}, \quad \sec = \frac{1}{\cos}, \quad \csc = \frac{1}{\sin}.$$

Uniqueness

We show first that at most one twice-differentiable function satisfies each of (13.2) and (13.3). A key observation is that the differential equation $y'' + y = 0$ is *linear*, in the sense that if f and g are solutions and c is a constant, then $(cf + g)$ is also a solution. Most differential equations do not have this property.

Proposition 13.2. *Let $y : \mathbf{R} \to \mathbf{R}$ be a twice-differentiable function satisfying $y'' + y = 0$ on $\mathbf{R}$. If $y(0) = y'(0) = 0$, then $y(x) = 0$ for all x.*

Proof. If $y'' + y = 0$ on $\mathbf{R}$, then we deduce that

$$\begin{aligned} \big((y')^2 + y^2\big)' &= 2y'y'' + 2yy' && \text{the chain rule} \\ &= 2y' \cdot (y'' + y) && \text{factoring} \\ &= 0 && \text{by hypothesis.} \end{aligned}$$

This means that the function $(y')^2 + y^2$ is constant on $\mathbf{R}$. Evaluating at 0 and using $y'(0) = y(0) = 0$, we find that $(y')^2 + y^2 = 0$ on $\mathbf{R}$, which in turn implies that y vanishes identically. □

Corollary 13.3. *Let $f : \mathbf{R} \to \mathbf{R}$ be a function satisfying $f'' + f = 0$, $f(0) = a$ and $f'(0) = b$. Then $f(x) = a\cos x + b\sin x$ for all $x \in \mathbf{R}$.*

Proof. Set $y = f - (a\cos + b\sin)$. Linearity of the differential equation $f'' + f = 0$ implies y is also a solution. The initial conditions on sin and cos imply that $y(0) = y'(0) = 0$. By Proposition 13.2, y is the zero function. □

It is difficult to overestimate the importance of this corollary. The basic properties of the trig functions are all immediate consequences, obtained by cooking up functions and showing that they satisfy the defining differential equation with appropriate initial conditions.

Existence

At present we have no logical basis for believing the differential equation $y'' + y = 0$ has any non-trivial solutions at all. We *do*, however, possess a powerful tool to attempt to *guess* the form of a solution, namely power series. Let us assume that

$$y(x) = \sum_{k=0}^{\infty} a_k x^k = a_0 + a_1 x + a_2 x^2 + \cdots$$

is a real-analytic solution of (13.2). Using the differential equation, we deduce the coefficients. Term-by-term differentiation (and shifting the index of summation) gives

$$\begin{aligned} y'(x) &= \sum_{k=0}^{\infty} (k+1) a_{k+1} x^k, \\ y''(x) &= \sum_{k=0}^{\infty} (k+2)(k+1) a_{k+2} x^k. \end{aligned} \tag{13.4}$$

The initial conditions determine the first two coefficients: $y(0) = a_0 = 0$, and $y'(0) = a_1 = 1$. Because $y'' = -y$ by assumption, equation (13.4) implies

$$a_{k+2} = -\frac{a_k}{(k+2)(k+1)} \qquad \text{for} \quad k \geq 0. \tag{13.5}$$

We find immediately that $0 = a_0 = a_2 = a_4 = \cdots$, while $a_3 = -1/(3{\cdot}2)$, $a_5 = 1/(5 \cdot 4 \cdot 3 \cdot 2)$, and so on. With a bit of thought, we guess that

$$a_{2k} = 0, \qquad a_{2k+1} = \frac{(-1)^k}{(2k+1)!} \qquad \text{for} \quad k \geq 0,$$

which is easily proven by induction on k. Thus

$$y(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots \tag{13.6}$$

is a *candidate* solution of (13.2). You should quickly check formally that the second derivative of this series is the negative of the original. This argument proves merely that *if* (13.2) has a real-analytic solution, *then* this solution is given by (13.6). Now we verify that (13.6) is indeed a real-analytic solution of (13.2). The series in (13.6) converges provided the ratio of consecutive non-zero terms approaches a limit that is smaller than 1 in absolute value. However, for all $x \in \mathbf{R}$,

$$\lim_{k\to\infty} \left| \frac{a_{2k+3}\, x^{2k+3}}{a_{2k+1}\, x^{2k+1}} \right| = \lim_{k\to\infty} \left| \frac{(2k+1)!\, x^2}{(2k+3)!} \right| = \lim_{k\to\infty} \left| \frac{x^2}{(2k+3)(2k+2)} \right| = 0,$$

so the power series (13.6) converges absolutely for all real x, and therefore defines a function $s : \mathbf{R} \to \mathbf{R}$. Termwise differentiation shows that $s'' + s = 0$ on $\mathbf{R}$; the choice of coefficients was motivated by the wish that this equation hold, after all! The initial values $s(0) = 0$ and $s'(0) = 1$ were also built into the choice of coefficients; we have therefore shown that (13.2) *has* a solution, in fact, has a real-analytic solution.

Equation (13.3) may be treated by parallel arguments, see Exercise 13.1. We shall henceforth use the fact that (13.2) and (13.3) have real-analytic solutions, and that the respective power series converge on all of $\mathbf{R}$.

Summary

By a combination of judicious guessing and appropriate use of powerful tools, we have shown there exist real-analytic functions sin and cos : $\mathbf{R} \to \mathbf{R}$ that satisfy

$$\begin{aligned} \sin'' &= -\sin, & \sin 0 &= 0, & \sin' 0 &= 1 \\ \cos'' &= -\cos, & \cos 0 &= 1, & \cos' 0 &= 0 \end{aligned}$$

These functions are defined on $\mathbf{R}$ by the power series

$$\sin x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!}, \qquad \cos x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k)!}. \tag{13.7}$$

Further, *every* twice differentiable function $y : \mathbf{R} \to \mathbf{R}$ that satisfies the differential equation $y'' + y = 0$ is a linear combination of sin and cos, and is consequently real-analytic. Finally, sin and cos are *characterized* by the initial-value problems they satisfy. In order to show that some function f is the sine function, it suffices to show that $f'' + f = 0$, and that $f(0) = 0$, $f'(0) = 1$.

Several useful properties of sin and cos can be derived from this characterization. These are collected as Theorem 13.4 below. This characterization will also be used to relate geometric definitions with our analytic definitions; we will define functions in terms of areas of angular sectors and prove (geometrically) that these functions satisfy the initial-value problems that characterize the sine and cosine functions.

Theorem 13.4. *The sine function is odd; the cosine function is even. The derivatives of* sin *and* cos *are given by*

$$\sin' = \cos, \qquad \cos' = -\sin. \tag{13.8}$$

For all $x \in \mathbf{R}$, $\sin^2 x + \cos^2 x = 1$. *Finally,* sin *and* cos *satisfy the following addition formulas:*

(13.9)
$$\begin{aligned} \sin(a+b) &= \sin a \cos b + \sin b \cos a \\ \cos(a+b) &= \cos a \cos b - \sin a \sin b \end{aligned} \qquad \text{for all } a, b \in \mathbf{R}.$$

In particular, $\sin(2x) = 2 \sin x \cos x$ *and* $\cos(2x) = \cos^2 x - \sin^2 x$ *for all* $x \in \mathbf{R}$.

Proof. It is apparent that sin is an odd function from its power series representation; however, a direct proof (in the spirit of the theorem) can be given using Corollary 13.3. Indeed the function $g : \mathbf{R} \to \mathbf{R}$ defined by $g(x) = \sin(-x)$ satisfies the differential equation $g'' + g = 0$ and the initial conditions $g(0) = 0$, $g'(0) = -1$, so the corollary implies $g = -\sin$. Evenness of cos is seen similarly.

If $y = \sin'$, then $y'' + y = 0$; this follows immediately upon differentiating the equation $\sin'' + \sin = 0$. But $y(0) = 1$ by definition of sin, and $y'(0) = \sin'' 0 = -\sin 0 = 0$. By Corollary 13.3, $\sin' = \cos$. A similar argument shows $\cos' = -\sin$.

Consider the function $f = \sin^2 + \cos^2$. The results of the previous paragraph imply that

$$f' = 2 \sin \sin' + 2 \cos \cos' = 2 \sin \cos + 2 \cos(-\sin) = 0,$$

which means f is constant. Since $f(0) = \sin^2 0 + \cos^2 0 = 1$, f is equal to 1 everywhere.

To prove the addition formulas, fix $b \in \mathbf{R}$ and consider the function y defined by $y(x) = \sin(x + b)$. The chain rule implies $y'' + y = 0$ on $\mathbf{R}$, and the derivative formula for sin implies $y'(x) = \cos(x + b)$ for all $x \in \mathbf{R}$. Substituting $x = 0$ gives $y(0) = \sin b$ and $y'(0) = \cos b$, so

$$\sin(a + b) = y(a) = \sin a \cos b + \sin b \cos a \qquad \text{for all } a \in \mathbf{R}$$

by Corollary 13.3. The addition formula for cos is proved similarly. □

Some standard limits are easy consequences of the power series representations of sin and cos. These limits are usually derived from geometric considerations and used to prove the derivative formulas in Theorem 13.4.

$$\lim_{x \to 0} \frac{\sin x}{x} = 1; \qquad \lim_{x \to 0} \frac{1 - \cos x}{x^2} = \frac{1}{2}.$$

Though each limit can be derived easily from l'Hôpital's rule, it is not logically permissible to do so if one plans to use the result to deduce the formula $\sin' = \cos$, since the resulting argument would be circular! In any case, the power series for sin and cos allow these limits to be evaluated directly. For $x \neq 0$,

$$\frac{\sin x}{x} = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{(2k+1)!} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \cdots.$$

By the ratio test, the series on the right represents a continuous function on $\mathbf{R}$, hence may be evaluated at 0 by setting $x = 0$; plainly this gives 1. The second limit is treated in Exercise 13.3.

Periodicity

In this section we prove that sin and cos are periodic. Their common period will be *defined* to be 2π; this is a non-geometric definition of the fundamental constant π, in contrast to the usual geometric definition, such as "the area of a unit disk" or "one-half the perimeter of a unit circle." The present definition is amenable to theoretical purposes and to numerical evaluation. Naturally, the geometric definitions will be recovered as theorems.

A physicist would suspect that sin and cos exhibit oscillatory behavior on two grounds: First, the equation $y''+y=0$ is the equation of motion for a mass on a spring in suitable units. (The fact that $(y')^2+y^2$ is constant is, in this situation, exactly conservation of energy.) Second, the equation $y''=-y$ says qualitatively that when y is positive, its graph is concave down, and *vice versa*. Thus the graph of y always bends toward the horizontal axis. Since the equation is "time independent", each time the solution crosses the axis from below to above, the solution is in the same "physical state" as it was the last time it crossed in this direction, so its future behavior repeats its past behavior.

In a sense, the mathematical proof of periodicity simply makes the physical intuition precise. The first step is to prove that the cosine function has a smallest positive zero, which for the moment we shall denote by α. Existence of α is accomplished via the following estimate, whose proof is deferred to the end of this section for conceptual continuity.

Proposition 13.5. *For all real x,* $1-\dfrac{x^2}{2!}\le\cos x\le 1-\dfrac{x^2}{2!}+\dfrac{x^4}{4!}$.

Granting this result, we find that

$$0\le\cos\sqrt{2},\qquad \cos\Big(\sqrt{6-\sqrt{12}}\Big)\le 0,$$

since $\sqrt{6-\sqrt{12}}$ is the first positive root of the quartic upper bound; see Figure 13.1. The intermediate value theorem implies cos has a zero between $\sqrt{2}\simeq 1.41421$ and $\sqrt{6-\sqrt{12}}\simeq 1.59245$. We now define $\pi=2\alpha$, and observe that Proposition 13.5 implies

$$2.82842\le\pi\le 3.1849.$$

These crude bounds are analogous to the estimate $2\le e\le 4$ that was obtained immediately from the definition of e. We will eventually find numerical series that converge fairly rapidly to π, which allows more accurate bounds to be obtained.

Returning to the main argument, Proposition 13.5 implies that cos has a smallest positive root: $\cos\alpha=0$, and $\cos'\alpha=-\sin\alpha$ is either 1 or -1 because $\cos^2+\sin^2=1$. Because $\cos 0=1$ and α is the *smallest* positive zero, cos is non-negative on the interval $[0,\alpha]$. It follows that $\cos'\alpha=-1$, for if it were 1 then cosine would be negative on some interval to the left of α. Evenness of cos implies that the largest negative zero of cos is $-\alpha$; in particular, there is an interval of length π, namely $(-\alpha,\alpha)$, on which cos is positive.

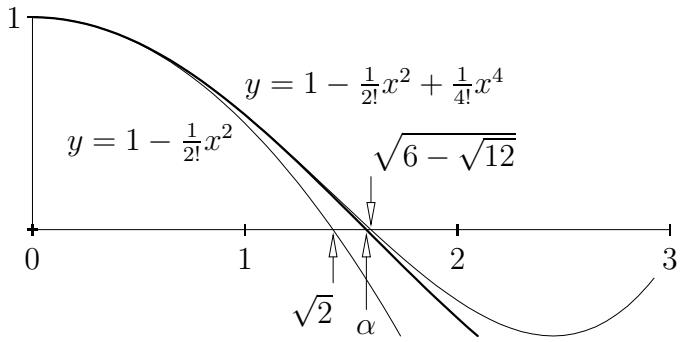


Figure 13.1: The smallest positive zero of the cosine function (bold).

By the addition formula for sin,

$$\begin{aligned}\sin(x+\alpha) &= \sin x \cos\alpha + \sin\alpha \cos x \\ &= \sin\alpha \cos x = \cos x \quad \text{for all } x \in \mathbf{R},\end{aligned}$$

since $\sin\alpha = -\cos'\alpha = 1$. Geometrically, the graph of cos is the graph of sin translated to the left by α. A similar argument shows that $\sin(x+\pi) = -\sin x$ for all $x \in \mathbf{R}$. Applying this equation twice shows that

$$\sin(x+2\pi) = -\sin(x+\pi) = \sin x \qquad \text{for all } x \in \mathbf{R}. \tag{13.10}$$

The cosine function is 2π-periodic as well since $\cos x = \sin(x+\alpha)$ for all $x \in \mathbf{R}$. Finally, there is no smaller positive period, since cos is positive on $(-\alpha, \alpha)$ and negative on $(\alpha, 3\alpha)$. In other words, the fact that 2π is the smallest positive period of sin and cos is a consequence of the fact that α is the smallest positive zero of cos.

The remaining piece of the proof of periodicity is Proposition 13.5. The argument is nothing more than repeated integration of an elementary inequality, but is completely different in character than the arguments above. To start, note that the equation $\sin^2 + \cos^2 = 1$ implies that $-1 \le \cos t \le 1$ for all $t \in \mathbf{R}$. Fix $x > 0$ and integrate from 0 to x, using the fundamental theorem of calculus:

$$-x \le \int_0^x \cos t \, dt = \sin t \Big|_{t=0}^{x} = \sin x \le x.$$

Thus $-t \le \sin t \le t$ for $t \ge 0$. Integrating *this* from 0 to x gives

$$-\frac{x^2}{2} \le \int_0^x \sin t \, dt = 1 - \cos x \le \frac{x^2}{2},$$

and since $x \geq 0$ was arbitrary it follows that

$$1 - \frac{t^2}{2} \leq \cos t \leq 1 \qquad \text{for } t \geq 0.$$

Another integration (and renaming of variable) gives $x - (x^3/6) \leq \sin x \leq x$ for $x \geq 0$, and a fourth gives $x^2/2 - x^4/24 \leq 1 - \cos x \leq x^2/2$, or

$$1 - \frac{x^2}{2} \leq \cos x \leq 1 - \frac{x^2}{2} + \frac{x^4}{24} \quad \text{for } x \geq 0. \tag{13.11}$$

This last set of inequalities is also true for $x < 0$ because each term is even in x. This completes the proof of Proposition 13.5.

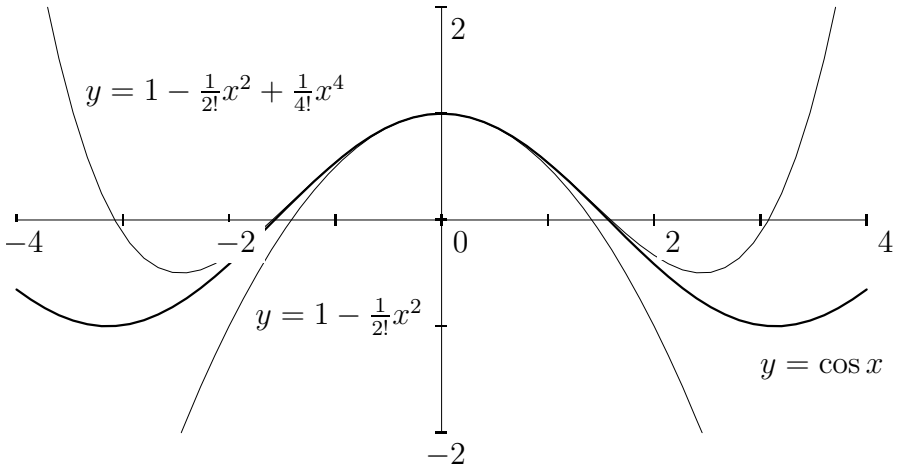


Figure 13.2: Upper and lower bounds on the cosine function.

The power series for cos has alternately positive and negative terms, and equation (13.11) suggests that the odd partial sums (ending with a positive term) are all upper bounds of cos while the even partial sums (ending with a negative term) are all lower bounds. The visual evidence is compelling, Figure 13.2. The claims just outlined are indeed true, as can be shown by induction on the process in the proof of Proposition 13.5. Moreover, the approximations get better as the degrees of the approximating polynomials get larger. It is important to emphasize, however, that the conclusion *cannot* be deduced solely on the basis of the signs of the terms; it is essential to consider the actual coefficients in the power series. Polynomial approximation is investigated systematically in Chapter 14.

13.2 Auxiliary Trig Functions

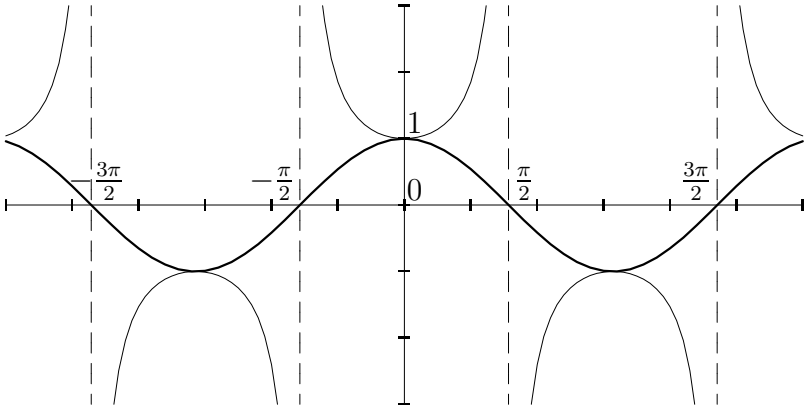


Figure 13.3: The graphs of cos (bold) and sec.

The sine function vanishes exactly at integer multiples of π, while the cosine function vanishes at "odd half-integer" multiples of π, namely at $(k + \frac{1}{2})\pi$ for $k \in \mathbf{Z}$. Both sin and cos are "anti-periodic" with period π in the sense that

$$\sin(x + \pi) = -\sin x, \quad \text{and} \quad \cos(x + \pi) = -\cos x \qquad \text{for all } x \in \mathbf{R}.$$

The secant function, $\sec = 1/\cos$, is undefined where cos vanishes, is even and has period 2π, and is anti-periodic with period π. Because $|\cos x| \leq 1$ for all $x \in \mathbf{R}$, $|\sec x| \geq 1$ for all real x in the domain of sec. The cosecant function, $\csc = 1/\sin$, satisfies $\csc(x + \frac{\pi}{2}) = \sec x$ because of the analogous relation between sin and cos.

The tangent function, $\tan = \sin / \cos$, is undefined where cos vanishes, and is periodic with period π (why?). The tangent function is odd, as a quotient of an odd function by an even function. Its behavior is completely determined by its behavior on the fundamental interval $(-\pi/2, \pi/2)$, to which we now turn.

The tangent function is differentiable on $(-\pi/2, \pi/2)$, and its derivative is found by the quotient rule to be

$$\tan' = \frac{\cos \sin' - \sin \cos'}{\cos^2} = \frac{\cos^2 + \sin^2}{\cos^2} = \frac{1}{\cos^2} = \sec^2 .$$

In particular, $\tan' > 0$ on $(-\pi/2, \pi/2)$, which implies tan is increasing on this interval. As already noted, cos is positive on $(-\pi/2, \pi/2)$ while

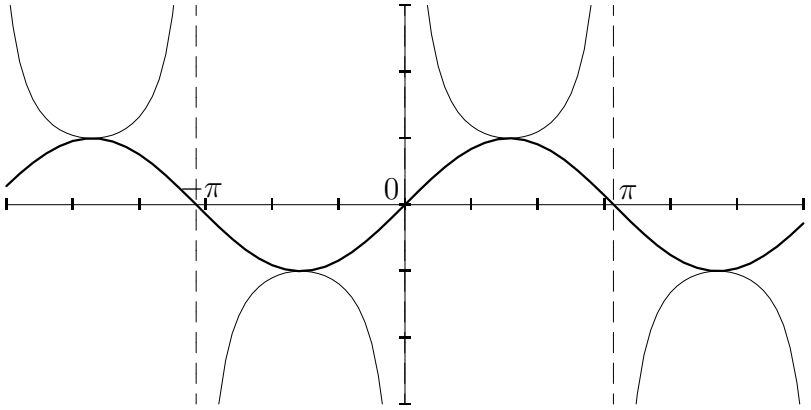


Figure 13.4: The graphs of sin (bold) and csc.

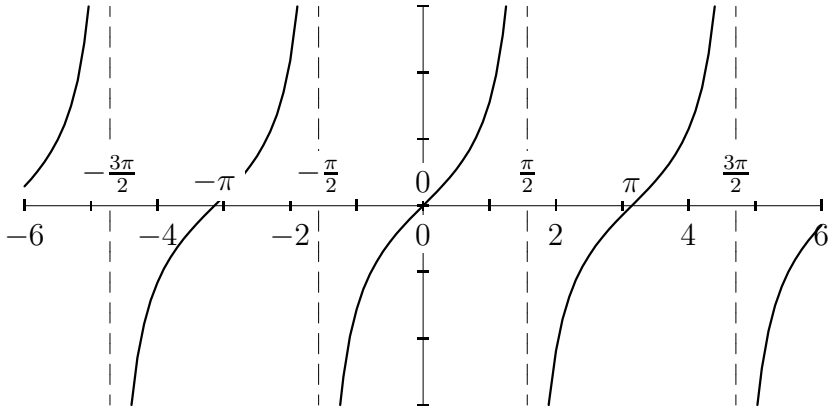


Figure 13.5: The graph of tan.

sin is positive on $(0, \pi/2)$ and thus negative on $(-\pi/2, 0)$. As $x \to \pi$ from below, $\tan x \to +\infty$, and since tan is odd,

$$\lim_{x \to -\pi/2^+} \tan x = -\infty.$$

In summary, tan maps $(-\pi/2, \pi/2)$ bijectively to $\mathbf{R}$, and is increasing on every interval of the form $\big((k - \frac{1}{2})\pi, (k + \frac{1}{2})\pi\big)$, with $k \in \mathbf{Z}$.

The cotangent function, cot = cos / sin, is not exactly the reciprocal of tan because of zeros and poles (places where the denominator vanishes); however, these functions *are* reciprocal everywhere they are both defined, zeros of tan are exactly poles of cot, and *vice versa*. The

derivative of cot is found to be $-1/\csc^2$, which shows that cot is decreasing on every interval in its domain, in particular on every interval of the form $\big(k\pi, (k+1)\pi\big)$ with $k \in \mathbf{Z}$.

Hyperbolic Trig Functions

The six trigonometric functions mentioned so far are sometimes called *circular* trig functions, because of their connection with the geometry of circles. Indeed, the identity $\cos^2 + \sin^2 = 1$ means that the point $(\cos t, \sin t)$ lies on the unit circle for all real t. There is a "dual" family of functions called *hyperbolic* trig functions, that have analogous names with an "h" appended, as in cosh, sinh (variously pronounced "cinch" or "shine"), and tanh (rhymes with "ranch"). These are, perhaps cryptically, defined directly in terms of the natural exponential function. The hyperbolic cosine and sine functions are the even and odd parts of exp, respectively:

$$\cosh x = \frac{e^x + e^{-x}}{2}, \qquad \sinh x = \frac{e^x - e^{-x}}{2}. \tag{13.12}$$

The auxiliary hyperbolic functions are defined by analogous equations,

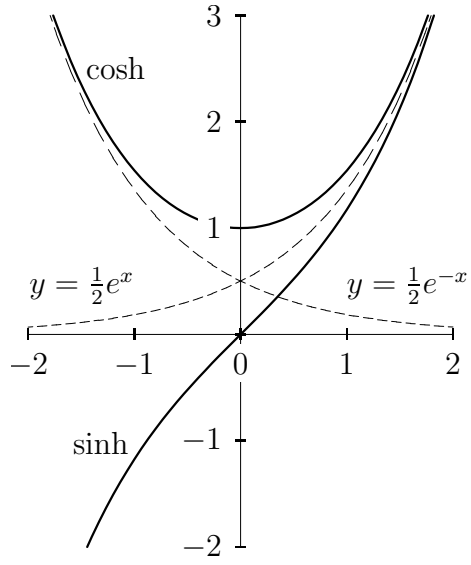


Figure 13.6: The graphs of cosh and sinh.

e.g.

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \qquad \operatorname{sech} x = \frac{1}{\cosh x}.$$

The remaining functions, coth and csch, are rarely encountered, but

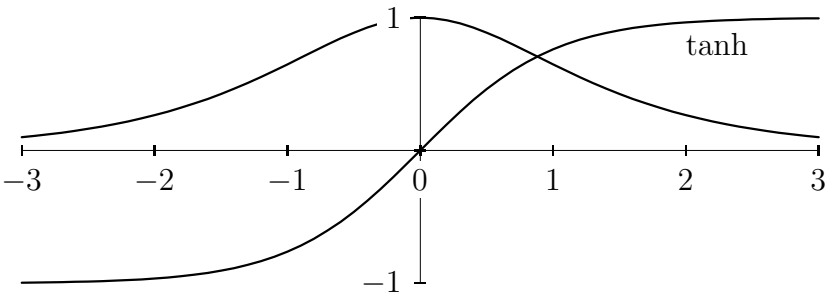


Figure 13.7: The graphs of tanh and sech.

the first four arise surprisingly often, in settings as diverse as hanging chains, soap films, non-Euclidean geometry, and solitary waves. There are numerous formal similarities between the circular and hyperbolic trig functions, some of which are investigated below. The underlying reason for these similarities is both deep and simple, but cannot be seen without defining all the functions over the set of complex numbers, see Chapter 15.

Simple calculations (left as exercises) verify that $\cosh^2 - \sinh^2 = 1$, and that

$$\cosh' = \sinh, \quad \sinh' = \cosh, \quad \tanh' = \operatorname{sech}^2. \tag{13.13}$$

The equation $\cosh^2 - \sinh^2 = 1$ means that the point $(\cosh t, \sinh t)$ lies on the unit hyperbola (with Cartesian equation $x^2 - y^2 = 1$) for all real t. The derivative expressions are analogous to circular trig formulas, but contain no signs, and can be traced to the fact that sinh and cosh are characterized as solutions of differential equations:

$$\begin{aligned} \sinh: &\quad y'' - y = 0, &\quad y(0) = 0, &\quad y'(0) = 1, \\ \cosh: &\quad y'' - y = 0, &\quad y(0) = 1, &\quad y'(0) = 0. \end{aligned}$$

There are addition formulas for sinh and cosh analogous to (13.9), as you can check directly with a bit of perseverance. To complete the

analogy, the power series representations of sinh and cosh may be found from the power series for exp; the result,

$$\sinh x = \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)!} \qquad \cosh x = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!},$$

clearly shows the similarity between the circular and hyperbolic trig functions.

13.3 Inverse Trig Functions

Each circular trigonometric function is periodic, hence has no inverse. Among the hyperbolic trig functions, sinh and tanh are injective, hence have "global" inverses. The functions cosh and sech are even, hence not one-to-one, but each is injective when restricted to the positive real axis. In this section we will investigate branches of inverse of the various trig functions. Perhaps the most remarkable feature is that while the inverse functions are not algebraic functions, their *derivatives* are algebraic. This is no accident, but a straightforward consequence of the differential equations that characterize the elementary trig functions.

The Functions arcsin and arccos

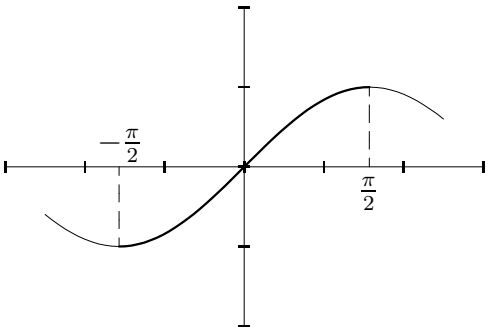


Figure 13.8: Sin.

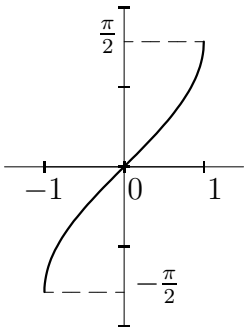


Figure 13.9: arcsin.

Cosine is positive on the interval $(-\pi/2, \pi/2)$; thus the sine function is *increasing* on this interval, since $\sin' = \cos$. Because $\sin(\pi - x) = \sin x$ for all real x, there is no larger open interval on which sin is injective. The restriction of sin to the closed interval $[-\pi/2, \pi/2]$ is denoted Sin.

The inverse function $\mathrm{Sin}^{-1} : [-1,1] \to [-\pi/2, \pi/2]$, sometimes denoted arcsin, is called the *principle branch of arcsine*. Thus

$$\begin{array}{rcll} \sin(\mathrm{Sin}^{-1}x) & = & x & \text{for all } x \in [-1,1], \\ \mathrm{Sin}^{-1}(\sin x) & = & x & \text{for all } x \in [-\pi/2, \pi/2]. \end{array} \tag{13.14}$$

The sine function is *decreasing* on $[\pi/2, 3\pi/2]$ because $\sin(\pi - x) = \sin x$. Periodicity implies that sin is one-to-one on $\left[(k-\frac{1}{2})\pi, (k+\frac{1}{2})\pi\right]$ for every integer k. For each k, there is a corresponding branch of arcsin, namely the inverse of the restriction of sin to $\left[(k-\frac{1}{2})\pi, (k+\frac{1}{2})\pi\right]$. On rare occasions when one considers a non-principle branch of arcsin, it is denoted $\sin^{-1}$, and k is supplied by context.

More-or-less identical remarks hold for cos. The principle branch is $\mathrm{Cos}^{-1} : [-1,1] \to [0,\pi]$, and for each integer k there is a branch of arccos taking values in $[k\pi, (k+1)\pi]$. The identity $\sin(x + \frac{\pi}{2}) = \cos x$ becomes

$$\mathrm{Cos}^{-1}x = \frac{\pi}{2} + \mathrm{Sin}^{-1}x \qquad \text{for all } x \in [-1,1]. \tag{13.15}$$

We now wish to find the derivative of Sin^{-1}. First of all, $\sin' = \cos$ is non-vanishing on $(-\pi/2, \pi, 2)$, so Sin^{-1} is differentiable on $(-1,1)$. This means we may differentiate the first equation in (13.14):

$$\cos(\mathrm{Sin}^{-1}x) \cdot (\mathrm{Sin}^{-1})'(x) = 1 \qquad \text{for all } x \in (-1,1).$$

The function Sin^{-1} takes values in $(-\pi/2, \pi, 2)$, and cos is positive on this interval. Thus $\cos = \sqrt{1-\sin^2}$ on this interval, so the previous equation can be rewritten

$$\begin{aligned} (\mathrm{Sin}^{-1})'(x) &= \frac{1}{\cos(\mathrm{Sin}^{-1}x)} = \frac{1}{\sqrt{1-\sin^2(\mathrm{Sin}^{-1}x)}} \\ &= \frac{1}{\sqrt{1-x^2}} \qquad \text{for all } x \in (-1,1). \end{aligned} \tag{13.16}$$

Because Sin^{-1} and Cos^{-1} differ by an additive constant, their derivatives are equal:

$$(\mathrm{Cos}^{-1})'(x) = \frac{1}{\sqrt{1-x^2}} \qquad \text{for all } x \in (-1,1). \tag{13.17}$$

These equations will be crucial when we equate geometric definitions of sin and cos with our analytic definitions.

The Other Circular Trig Functions

The tangent function maps $(-\pi/2, \pi/2)$ bijectively to $\mathbf{R}$. The inverse of the restriction of tan to this interval is the *principle branch of arctan*, denoted $\mathrm{Tan}^{-1} : \mathbf{R} \to (-\pi/2, \pi/2)$. Because tan is π-periodic, the other branches of arctan differ from the principle branch by an added multiple of π. The derivative of Tan^{-1} is found by differentiating the first of

$$\begin{array}{rcll} \tan(\mathrm{Tan}^{-1} x) & = & x & \text{for all } x \in \mathbf{R}, \\ \mathrm{Tan}^{-1}(\tan x) & = & x & \text{for all } x \in [-\pi/2, \pi/2]. \end{array} \tag{13.18}$$

The short calculation is left as an exercise; the result is

$$(\mathrm{Tan}^{-1})'(x) = \frac{1}{1+x^2} \qquad \text{for all } x \in \mathbf{R}. \tag{13.19}$$

This is even more remarkable than equation (13.16); the derivative of Tan^{-1} is a *rational* function, not merely an algebraic function! To emphasize a philosophical point, the derivative of a rational function is always a rational function, but an antiderivative need not be. We have already seen this for the reciprocal function, but the point bears repeating.

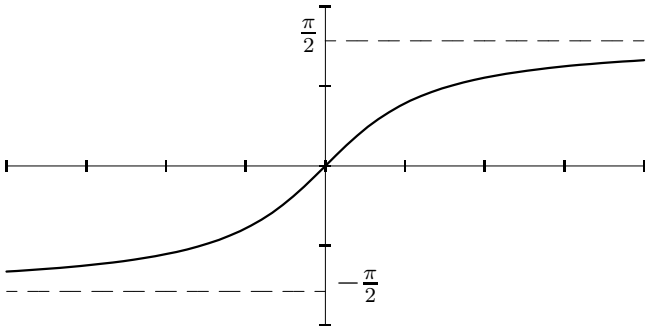


Figure 13.10: The principle branch of arctan.

The inverses of the other circular trig functions are less prominent in applications, though arcsec does arise in evaluating certain integrals. To describe the domain and image in detail, consider the cosine function restricted to $[0, \pi]$. Its reciprocal, the restriction of sec, is defined on the union $[0, \pi/2) \cup (\pi/2, \pi]$. The principle branch of arcsec is the inverse of this restriction; its domain is $(-\infty, -1] \cup [1, \infty)$.

The Hyperbolic Trig Functions

The inverse hyperbolic trig functions can be calculated directly from their definitions. To solve the equation $y = \cosh x = (e^x + e^{-x})/2$ for x, multiply both sides by $2e^x$ and rearrange to get

$$(e^x)^2 - (2y)e^x + 1 = 0.$$

This is a quadratic equation in e^x, and can be solved using the quadratic formula:

$$x = \log\bigl(y \pm \sqrt{y^2 - 1}\bigr), \qquad y \geq 1. \tag{13.20}$$

We expect two real branches because cosh is not one-to-one. As a consistency check, observe that $y - \sqrt{y^2-1} = 1/(y + \sqrt{y^2-1})$ for $|y| \geq 1$, and both these quantities are positive for $y \geq 1$, so

$$\log\bigl(y \pm \sqrt{y^2 - 1}\bigr) = \mp \log\bigl(y - \sqrt{y^2 - 1}\bigr),$$

and the two branches do indeed differ by a sign. A similar calculation shows that $\sinh^{-1}$ is defined by

$$\sinh^{-1} x = \log\bigl(x + \sqrt{x^2 + 1}\bigr). \tag{13.21}$$

There is no ambiguity with signs because only this choice leads to a real-valued function when x is real. You may also verify that the expression on the right is an odd function of x.

The inverse of tanh is even easier to find. Simple algebra shows that

$$y = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} - 1}$$

if and only if

$$\tanh^{-1} y = x = \frac{1}{2} \log\left(\frac{1-y}{1+y}\right), \qquad -1 < y < 1. \tag{13.22}$$

As a consistency check, the expression on the right is an odd function of y.

The derivatives of these inverse functions are algebraic functions that look very similar to their circular counterparts. It is left as an

exercise to show that

$$
\begin{aligned}
(\cosh^{-1})'(x) &= \frac{1}{\sqrt{x^2-1}}, \\
(\sinh^{-1})'(x) &= \frac{1}{\sqrt{x^2+1}}, \\
(\tanh^{-1})'(x) &= \frac{1}{x^2-1}.
\end{aligned}
\tag{13.23}
$$

13.4 Geometric Definitions

This section does not contribute, strictly speaking, to the logical development of the circular trigonometric functions. The intent is rather to connect the trig functions as already introduced with the familiar pictures of arclength along the unit circle and area enclosed by a circular sector. The presentation is relatively informal, and uses pictures and geometric intuition freely. In order to emphasize that something non-trivial is being shown, geometric versions of trig functions will be denoted with capital letters (e.g., COS) until they are proven to be the same as the functions defined analytically above.

The word "trigonometry" comes from Greek roots meaning "triangle measurement." It is a feature of Euclidean geometry that similar triangles exist; there exist non-congruent triangles that have the same internal angles.[1] The shape of a *right* triangle is determined, up to similarity, by the ratios of its side lengths. It is also determined, up to similarity, by *one* of its acute angles. The circular trig functions are the ratios of side lengths as a function of an acute angle, see equation (13.24) below. Many students of trigonometry learn a mnemonic[2] of some sort to remember which function is which ratio. This definition is only sensible for acute angles; to define the trig functions for arbitrary real numbers one extends by symmetry and periodicity. In order to motivate these extensions, introduce a Cartesian coordinate system with the acute angle θ at the origin and the hypotenuse scaled

[1] Strange as it may seem, not all geometries have this property. Think of measuring patches of the surface of a sphere; the sides of triangles are arcs of great circles. If the three internal angles of a triangle are known, then the side lengths may be deduced; consequently, two triangles with the same internal angles are actually congruent.

[2] Like sohcahtoa.

to have unit length, as in the first half of Figure 13.11. The triangle itself is then demoted to a secondary role, and θ is allowed to be arbitrary, even negative (corresponding to a "clockwise" angle). The trigonometric ratios are defined to be

$$x = \mathrm{COS}\,\theta, \qquad y = \mathrm{SIN}\,\theta, \qquad \frac{y}{x} = \mathrm{TAN}\,\theta. \tag{13.24}$$

If an angle of 2Π corresponds to a full revolution, then SIN and COS are 2Π periodic. The angle θ is determined only up to an added multiple of 2Π by the point (x, y).

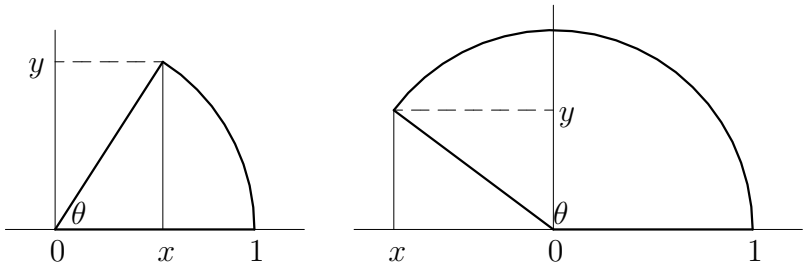


Figure 13.11: Circular sectors subtended by a ray through the origin.

It is perhaps intuitively clear that there is a numerical quantity called "angle," but it is probably not obvious how to measure it "naturally." For historical reasons originating in astronomy, the Babylonians divided the circle into 360 pieces, called *degrees*. Even modern English uses idioms based on this system.[3] There is nothing mathematically natural about degrees as a measure of angle, any more than base 10 notation is the natural means of writing integers. Instead, nature prefers[4] angles to be measured geometrically, either in terms of arclength around the circle, or in terms of areas of circular sectors.

The length of an arc of the unit circle is closely related to the area of the sector it defines. In Cartesian coordinates (u, v), the circle has equation $u^2 + v^2 = 1$. Each ray through the origin intersects the circle in a unique point (x, y). (The completeness axiom for $\mathbf{R}$ is implicit here.) Let θ be the arclength between $(0, 0)$ and (x, y), measured counterclockwise[5] along the circle, and let 2Π be the circumference. There is

[3]"No it doesn't," said the author, making a 180° reversal of his claim. "Wait, I was wrong. It does," he added, coming around a full 360.

[4]This claim will be fully justified shortly.

[5]This is also a convention, but a harmless one.

a corresponding sector of the circle, enclosed by the positive u-axis, the arc, and the ray, as in Figure 13.11. As was known to Archimedes, the area of the sector is $\theta/2$. He showed this with the following argument, see Figure 13.12 for the case where the entire disk is considered.

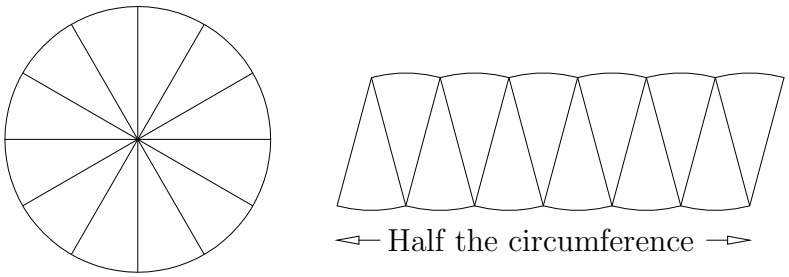


Figure 13.12: Archimedes' dissection of a disk to an approximate rectangle.

Let A be the area of the sector with angle θ at the origin. Divide the sector into N congruent pieces, each of which is approximately an isosceles triangle of base θ/N and height 1. Each slice has area approximately $\theta/2N$, so the total area A is approximately $\theta/2$. The approximation can be made arbitrarily accurate by taking N large, so $A = \theta/2$.[6] In particular, the area of the unit disk is Π.

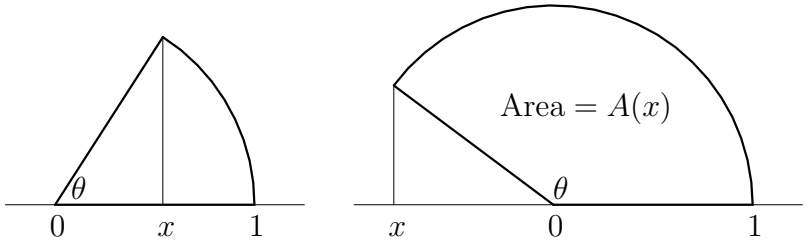
Intuitively, the sector can be cut into infinitely many infinitely thin triangles which are rearranged into a rectangle of height 1 and width $\theta/2$, but this intuition is not literally correct. The language of integrals is suitable for making this assertion rigorous, but you should once again be reminded of the Goethe quote at the beginning of the book.

Proposition 13.6. *The area of a circular sector pictured below is*

$$A(x) = \frac{x}{2}\sqrt{1-x^2} + \int_x^1 \sqrt{1-u^2}\,du$$

for $-1 \leq x \leq 1$.

[6] If two numbers are ε-close for all $\varepsilon > 0$, then they are equal.



Proof. Since (x, y) lies on the upper half of the unit circle, $y = \sqrt{1-x^2}$. If $x > 0$ (the left-hand picture), then $x\sqrt{1-x^2}/2$ is the area of the right triangle, while the integral is the area of the curved region. On the other hand, if $x < 0$ (the right-hand picture), then $x\sqrt{1-x^2}/2$ is negative, but has absolute value equal to the area of the right triangle, while the integral is the area of the entire enclosed region. Again, the sum—which is the *difference* in areas—is the area of the circular sector. □

The particular case $x = -1$ is interesting:

Corollary 13.7. $\dfrac{\Pi}{2} = \displaystyle\int_{-1}^{1} \sqrt{1-u^2}\, du.$

However, an even more substantial conclusion results from differentiating. By the fundamental theorem,

$$A'(x) = \frac{\sqrt{1-x^2}}{2} + x \cdot \frac{-x}{2\sqrt{1-x^2}} - \sqrt{1-x^2} = -\frac{1}{2\sqrt{1-x^2}}$$

for $-1 < x < 1$: The functions $2A$ and Cos^{-1} have the same derivative. In addition, they agree at 1, so *they are the same function.* This is the first link between the circular trig functions and the geometrically defined function A.

Note also that $A(-1) = \mathrm{Cos}^{-1}(-1) = \pi$; Corollary 13.7 implies that $\pi = \Pi$; the period of sin and cos is the circumference of the unit circle. To tie up the remaining loose end, Archimedes' theorem on areas of sectors says (in the notation of Figure 13.6) $x = \cos\theta$ for $\theta \in [0, \pi]$. This equation is true when $\theta \in [-\pi, 0]$ because of two facts: (i) cos is an even function, and (ii) the figure in Proposition 13.6 is symmetric on reflection in the horizontal axis, which exchanges θ and $-\theta$. Combining these observations, cos = COS on the interval $[-\pi, \pi]$, and since both functions are 2π-periodic, they are equal everywhere. This means the circular trig functions (defined as solutions of a differential equation) are the same as the functions COS and SIN defined as horizontal and

vertical coordinates of a point on a circle. The "variable" is measured not in degrees, but in *radians*—units of arclength along the circle.

It was asserted earlier that radians are the "natural" measure of angle. The main justification is that $\mathrm{SIN}' = \mathrm{COS}$ and $\mathrm{COS}' = -\mathrm{SIN}$. Suppose degrees had been used to define circular trig functions SIN° and COS°. (The function COS° is "just like COS, but takes input in degrees".) Equations like $\mathrm{COS}^\circ 90 = 0$ and $\mathrm{SIN}^\circ 90 = 1$ would hold (which would not be a problem), but the equations $(\mathrm{SIN}^\circ)' = \mathrm{COS}^\circ$ and $(\mathrm{COS}^\circ)' = -\mathrm{SIN}^\circ$ would be *false* (which would be badly inconvenient). To see what equations would replace them, observe that COS and COS° differ by scaling the domain. Precisely,

$$\mathrm{COS}\,(\pi \cdot \Theta/180) = \mathrm{COS}^\circ\Theta \qquad \text{for all } \Theta \in \mathbf{R},$$

since Θ degrees is the same angle as $\pi \cdot \Theta/180$ radians. From this equation, it is easy to check that

$$(\mathrm{COS}^\circ)' = -\frac{\pi}{180}\mathrm{SIN}^\circ.$$

This equation is at best aesthetically unpleasing, as it builds an arbitrary number (namely 180) into a fundamental trigonometric relation.

Exercises

Exercise 13.1 Mimic the construction of sin in detail to construct cos. The only difference is in the initial conditions. ◇

Exercise 13.2 Use termwise differentiation to give an alternative proof that $\sin' = \cos$ and $\cos' = -\sin$. ◇

Exercise 13.3 Prove that $\lim_{x\to 0} \dfrac{1-\cos x}{x^2} = \dfrac{1}{2}$, both with l'Hôpital's rule and using power series. ◇

Exercise 13.4 Evaluate $\lim_{x\to 0} \displaystyle\int_0^{x^2} \frac{1-cost}{x^6}\,dt$ ◇

Exercise 13.5 Prove that $\sec' = \sec\cdot\tan$ (don't forget to show the domains are the same). ◇

Exercise 13.6 Use the identity $\sec^2 = 1 + \tan^2$ to prove (13.19). ◇

Exercise 13.7 Establish the identities:

(a) $\tan(x+y) = \dfrac{\tan x + \tan y}{1 - \tan x \tan y}$

(b) $\cot 2x = \frac{1}{2}(\cot x - \tan x)$

For each, determine the set of x and y for which the identity holds. ◇

Exercise 13.8 Using the results of this chapter, evaluate the following:

(a) $\sin \frac{\pi}{4}$, $\cos \frac{\pi}{4}$, $\sec \frac{\pi}{4}$, $\tan \frac{\pi}{4}$.

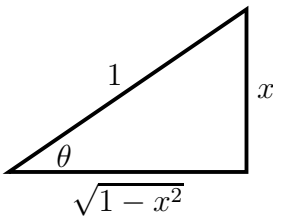
(b) $\sin \frac{\pi}{6}$, $\cos \frac{\pi}{6}$, $\sec \frac{\pi}{6}$, $\tan \frac{\pi}{6}$.

(c) $\sin \frac{\pi}{3}$, $\cos \frac{\pi}{3}$, $\sec \frac{\pi}{3}$, $\tan \frac{\pi}{3}$.

(d) $\sin \frac{\pi}{8}$, $\cos \frac{\pi}{8}$

Your answers should involve only square roots and rational numbers. ◇

Exercise 13.9 Let $0 < x < 1$, and let $\theta = \mathrm{Sin}^{-1} x$:



It follows immediately that $\cos\theta = \sqrt{1-x^2}$,

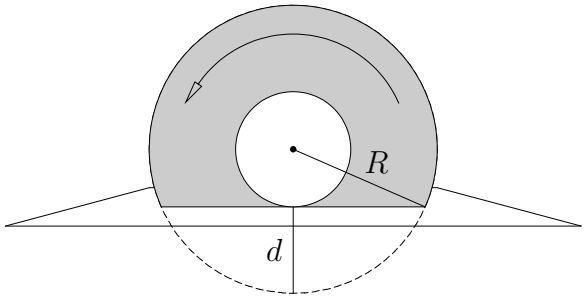
$$\sec\theta = \frac{1}{\sqrt{1-x^2}}, \quad \tan\theta = \frac{x}{\sqrt{1-x^2}}, \quad \cot\theta = \frac{\sqrt{1-x^2}}{x}.$$

Similarly, find $\cos \mathrm{Tan}^{-1} x$, $\sec \mathrm{Tan}^{-1} x$, and $\sin \mathrm{Tan}^{-1} x$. ◇

Exercise 13.10 Verify equation (13.13). ◇

Exercise 13.11 Verify equation (13.23). ◇

Exercise 13.12 As part of an industrial process, a thin circular plate of metal is spun about its axis while partially submerged in a vat of polymer. The horizontal view along the axis is shown:



If the wheel has radius R, find the depth d that maximizes the amount of polymer exposed to the air (shaded). ◇

The following multi-part exercise presents Ivan Niven's proof that π^2 is irrational. It follows that π itself is irrational.

Exercise 13.13 Define $f_n : [0,1] \to \mathbf{R}$ by

$$f_n(x) = \frac{x^n(1-x)^n}{n!}.$$

Prove each of the following assertions.

(a) If $0 < x < 1$, then $0 < f_n(x) < \frac{1}{n!}$.

(b) The derivatives $f_n^{(k)}(0)$ and $f_n^{(k)}(1)$ are integers for all $k \in \mathbf{N}$.

Assume $\pi^2 = p/q$ in lowest terms, and let

$$F_n(x) = q^n \sum_{k=0}^{n} (-1)^k f_n^{(2k)}(x) \pi^{2n-2k}.$$

(c) $F_n(0)$ and $F_n(1)$ are integers.

(d) $\pi^2 p^n f_n(x) \sin \pi x = \dfrac{d}{dx} \left(F_n'(x) \sin \pi x - \pi F_n(x) \cos \pi x \right)$.

(e) $F_n(1) + F_n(0) = \pi p^n \displaystyle\int_0^1 f_n(x) \sin \pi x \, dx$.

(f) Take $n \gg 1$ to deduce that $0 < F_n(0) + F_n(1) < 1$.

In short, if π^2 is rational, then there is an integer between 0 and 1. ◇

Chapter 14

Taylor Approximation

Armed with a collection of powerful mathematical tools (the mean value theorem, the fundamental theorems of calculus, and power series) and a rich library of examples (algebraic, exponential, logarithmic, and trigonometric functions), we turn to systematic investigation of calculational issues.

14.1 Numerical Approximation

You have probably learned from an early age that mathematics is "an exact science" and that problems have one right answer. Ironically, this impression is reinforced by electronic calculators, which evaluate many common functions to 8 decimal places (or more) at the push of a button. As mentioned in Chapter 2, no numerical answer returned by a calculator can be irrational, so most calculator results are only approximate. Sometimes lip service is paid to this fact by writing, for example, $e = 2.71828\ldots$ or $e \simeq 2.71828$, with little explanation of what the ellipsis or the squiggly equal sign mean. Some questions should come to mind immediately:

- How are constants like $\sqrt{2}$, e, and π defined if not as infinite decimals?
- What does it mean when a calculator returns a numerical value for a possibly irrational number?
- How does a calculator know the value to return? (Or "How did the person who programmed the calculator know?", to push things a step further back.)

You already know the answer to the first question (either that or it's time to re-read Chapters 5, 12, and 13!). The second question is answered by the A-notation of Chapter 2, which we review briefly. The third question occupies the remainder of the chapter.

When a calculator says $e = 2.71828$, it really means "The value of e rounded to five decimal places is 2.71828," which in turn means (by convention) that $2.718275 \leq e < 2.718285$, or (essentially) that

$$e = 2.71828 + A(0.5 \times 10^{-5}). \tag{14.1}$$

A rounded-off answer represents a limitation of knowledge; it asserts that the number in question lies within a certain interval of real numbers. Each additional decimal place corresponds to an interval that is one-tenth as long, so more decimals mean more information:

$$e = 2.718\,281\,828\,459\,045 + A(0.5 \times 10^{-15}).$$

Conversely, an approximation must be known about ten times more accurately to garner a single additional decimal place. It is not difficult to see why engineers and scientists are generally happy with 4 decimals of accuracy, and why 9 or 10 decimals are roughly the limits of measurement. For example, the distance to the moon is roughly 238,000 miles, or about 1.508×10^{10} inches. Using mirrors left by the Apollo astronauts to reflect lasers, scientists can measure the distance to the moon[1] to an accuracy of about 6 inches, which is just 9 decimals. An accuracy of 20 decimals would correspond to an experimental error on the order of one atomic diameter, and an accuracy of 50 decimals in this context is physically meaningless, because very small distances are not well-modeled by real numbers, but instead are subject to the consequences of quantum mechanics.

By contrast, a pure mathematician is unlikely to feel satisfied unless *arbitrarily many* decimals can be found; anything less is subject to uncertainty. A spectacular example is the number discovered by C. Hermite[2] in 1859, whose numerical value is

$$x_0 := e^{\pi\sqrt{163}} = 262,537,412,640,768,744 + A(10^{-12}).$$

Is x_0 exactly an integer? The numerical evidence is overwhelming: The error term is $\pm 0.000\,000\,000\,000\ldots$, so x_0 is an integer to one

[1] Really, the distance from a laser in a telescope to a mirror on the moon!

[2] air MEET

part in 10^{30}, an accuracy absolutely unattainable in scientific measurement. However, such "reasoning" is wishful thinking; indeed, such coincidences *must* happen in many situations. The error term is not zero, but $A(0.75\times10^{-12})$! Mathematicians are sometimes regarded as pedantic by experimental scientists ("They make you prove things that are obvious."), but mathematicians' skepticism is not gratuitous.

Even scientists and applied mathematicians have a vested interest in "mathematical precision", because numerical errors tend to grow rapidly in calculations, with a fixed number of decimals of accuracy lost at each multiplication. Many a student, when asked to evaluate $e^{10\log 2}$, will first round off the logarithm, $\log 2 \simeq 0.693$ (to 3 decimals), then multiply by 10 and exponentiate, obtaining $e^{10\log 2} \simeq 1022.494$. However, properties of exponentials and logarithms imply that the *exact value* is $2^{10} = 1024$. This example shows the loss of accuracy in a *single step*; even a simplified word problem may have three or four steps of this type, and careless or premature use of numerical constants can lead to a drastically wrong answer. Unfortunately, calculators have fostered the bad habit of plugging in numbers at the start of a calculation. A theoretical scientist or mathematician must be a fluent symbolic calculator, but even if you aspire to be successful as an experimental scientist, you must cultivate the ability to calculate symbolically.

These considerations explain mathematicians' insistence on procedural definitions of constants like e and π (the definitions are precise and can be turned into computational algorithms yielding arbitrary accuracy), rather than on numerical "specifications" (such as "$\pi \simeq 3.141\,592\,653\,589\,793\ldots$") which are not mathematical definitions at all.

To mathematicians, "evaluating" a numerical quantity usually means finding an expression that gives the exact value in terms of elementary functions and well-known transcendental constants. For example, the number π is defined to be $1/2$ the period of cos, while

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots$$

is defined to be the limit of the sequence of partial sums. We are unlikely to be able to evaluate a given infinite sum exactly, even if we can prove the sum exists. Of course, the sum can be evaluated with arbitrary numerical precision, which is partially satisfactory, but on this basis no mathematician would say the sum has been "evaluated." A very different state of affairs results if, say, the sum is shown to

be equal to $\pi^2/6$ (as it turns out to be in the above example), for in this case the exact value of the sum is known in terms of π, and π is as familiar and ubiquitous mathematically as an integer. Numerically such information is helpful because a numerical approximation of π allows the sum to be approximated without any hard work. Such gems as result from exact evaluation of a series or integral often shed light on a hidden phenomenon, and are therefore important beyond their intrinsic beauty.

14.2 Function Approximation

The discussion above carries over to functions. In analogy to equation (14.1), we hope to find expressions such as

$$e^x = \underbrace{\left(\sum_{k=0}^{n} \frac{x^k}{k!}\right)}_{\text{estimate}} + \underbrace{A\left(\frac{3|x|^{n+1}}{(n+1)!}\right)}_{\text{error}} \quad \text{for } |x| \leq 1. \tag{14.2}$$

In this equation, e^x is the number we wish to approximate. The first term on the right is our *estimate* of e^x, while the second term, the *error term*, is an upper bound on the difference between e^x and the estimate, and measures the accuracy of the estimate. The crucial feature of equation (14.2) is that the estimate and error are *polynomials* in x, which may be evaluated numerically using nothing but arithmetic operations. Taking $x = 1$ and $n = 6$ gives

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \frac{1}{6!} + A\left(\frac{3}{7!}\right),$$

or $e = 2.718 + A(0.0006)$. For greater accuracy, we would take a larger choice of n. An equation like (14.2) encapsulates infinitely many numerical conditions (one for each x), and is of obvious computational interest.

The aim of this section is to establish estimates analogous to (14.2) for functions that possess sufficiently many derivatives. The strategy is to approximate a function f near a point x_0 by choosing a *Taylor polynomial* p_n of degree at most n that "matches" f and its derivatives up to order n at x_0. When $n = 1$ this strategy yields the tangent line, a linear function whose derivative is $f'(x_0)$. Our aim is threefold:

- Give an effective computational procedure for finding p_n.

- Show that in an appropriate sense p_n is the "best approximating polynomial" of degree at most n.
- Find an effective computational bound on the error term.

Once these issues are resolved, the question, "How does a calculator know what numerical value to return?" is very nearly answered: The calculator is programmed with an algorithm for evaluating the Taylor polynomials of common functions, and returns a value that differs from the actual (mathematical) value by less than 0.5×10^{-8} (say). Taylor polynomials reduce the calculation of e^x or $\cos x$ to addition, subtraction, multiplication, and division. The arithmetic operations themselves are performed by the *floating point unit* (FPU), a circuit that operates at a level only slightly higher than the level at which we constructed the integers in Chapter 2.

Taylor Polynomials

Let f be a function that is defined on some open interval containing $x_0 \in \mathbf{R}$, and assume that f is N times differentiable for some positive integer N. We will construct a sequence $\{p_n\}_{n=0}^{N}$ of polynomials such that

(i) For each n, the polynomial p_n has degree at most n.

(ii) For $0 \leq k \leq n$, the derivatives $f^{(k)}(x_0)$ and $p_n^{(k)}(x_0)$ agree.

The sequence is uniquely specified by these two properties (as we shall see), and is called the sequence of *Taylor polynomials* of f at x_0 up to degree N. It turns out that the difference between p_n and p_{n-1} is a monomial of degree n, so p_N immediately determines every Taylor polynomial p_k with $k < N$. In practice, this means that we needn't write down the entire sequence, but only the polynomial p_N.

Remark 14.1 A Taylor polynomial p_n depends upon the function f, the "center" point x_0, and the degree n, but the function and center point are omitted to simplify the notation when they are clear from context. In applications, the center is usually 0, but it is theoretically useful to allow x_0 to be arbitrary. When necessary, the Taylor polynomial centered at x_0 is denoted p_{n,x_0}. □

To find the coefficients of a Taylor polynomial, we expand in powers of $(x - x_0)$ rather than in powers of x. Let f be n times differentiable

on a neighborhood of x_0, and write

$$p_n(x) = \sum_{j=0}^{n} a_j (x - x_0)^j$$

with $\{a_j\}_{j=0}^n$ unknown; the expression on the right is the general polynomial of degree at most n, as stipulated by condition (i). Next we equate coefficients, as in (ii).

Lemma 14.2. *With p_n as above, $p_n^{(k)}(x_0) = f^{(k)}(x_0)$ iff $a_k = \dfrac{f^{(k)}(x_0)}{k!}$.*

Proof. We compute $p_n^{(k)}(x_0)$ in terms of the coefficients of p_n by tallying up the contributions from the summands,

$$(*) \qquad \left.\frac{d^k}{dx^k}\right|_{x=x_0} a_j (x - x_0)^j.$$

If $j < k$, the derivative is identically zero (each differentiation lowers the exponent by 1), while if $j > k$ the kth derivative is divisible by $(x - x_0)$, and therefore vanishes at x_0. (See also Example 14.6.) The only nontrivial contribution to $(*)$ comes from the term with $j = k$; it is left to you to prove inductively that

$$(**) \qquad \left.\frac{d^k}{dx^k}\right|_{x=x_0} a_k (x - x_0)^k = k!\, a_k.$$

Thus $p_n^{(k)}(x_0) = k!\, a_k$. The lemma follows immediately. □

According to the lemma, Property (ii) is satisfied (the derivatives of p_n and f agree up to order n) iff

$$(14.3) \qquad p_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k.$$

This is the simple formula we seek. As long as we can calculate derivatives of f and evaluate them at x_0, we have an explicit representation of the Taylor polynomials.

The difference between consecutive Taylor polynomials is

$$p_n(x) - p_{n-1}(x) = \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n,$$

a monomial of degree n. This establishes the prior claim that each Taylor polynomial p_n "contains" all the polynomials p_k with $k < n$. To obtain p_k from p_n, simply drop all the terms of degree greater than k. To obtain p_n from p_{n-1}, we need only add a monomial (possibly zero) in degree n.

Remark 14.3 In Chapter 11, we saw that an arbitrary continuous function is uniformly approximated by polynomials on every closed, bounded interval. However, such an approximating sequence generally does not have the property that p_n and p_{n-1} differ by a monomial at all. What we gain in generality is paid for with loss of simplicity: Weierstrass polynomials require no differentiability assumptions, but knowledge of p_n does not tell us anything about p_k with $k < n$. □

Example 14.4 (The exponential function) Because $\exp^{(k)}(x) = \exp x$ for all positive integers k, $\exp^{(k)}(0) = 1$ for all k, so the Taylor polynomials of exp at 0 are

$$p_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}.$$

The Taylor polynomials are exactly the partial sums of the power series of exp, given by Theorem 12.7. This is a general feature of real-analytic functions, see Corollary 14.11. □

Example 14.5 (The circular trig functions) The elementary trig functions sin and cos have Taylor polynomials that are easy to calculate from the definition.[3] As for exp, the Taylor polynomials are truncations of the power series. For example, the degree $2n + 1$ Taylor polynomial of sin is

$$p_{2n+1}(x) = \sum_{k=0}^{n} (-1)^k \frac{x^{2k+1}}{(2k+1)!}.$$

In this example, the Taylor polynomial of degree $2n+2$ is equal to the Taylor polynomial of degree $2n+1$ because the even terms of the power series are zero. □

Example 14.6 (Power functions) Fix $a \in \mathbf{R}$ and a positive integer N, and consider the polynomial $f(x) = (x-a)^N$. Successive differentiation gives

$$f'(x) = N(x-a)^{N-1}$$
$$f''(x) = N(N-1)(x-a)^{N-2}$$

[3]The other trig functions do not, as it turns out.

and generally

$$\begin{aligned} f^{(k)}(x) &= N(N-1)\cdots(N-k+1)(x-a)^{N-k} \\ &= \frac{N!}{(N-k)!}(x-a)^{N-k} \qquad \text{for } k = 0, \ldots, N. \end{aligned}$$

Substituting into (14.3) gives the Taylor polynomial at $x_0 = 0$:

$$p_n(x) = \sum_{k=0}^{n} \binom{N}{k} (-a)^{N-k} x^k, \qquad n = 0, \ldots, N.$$

When $n = N$, the right-hand side is $(x-a)^N$ by the binomial theorem, and this remains true for $n \geq N$ because the higher derivatives of f vanish. It is no accident that the Nth-degree Taylor polynomial of the Nth degree polynomial f turned out to be f itself, see Theorem 14.9 below. □

Example 14.7 (The binomial series) Let α be a real number that is *not* a non-negative integer, and define $f(x) = (1+x)^\alpha$ for $x > -1$. The function f is infinitely differentiable at 0, and as in the previous example,

$$f^{(k)}(x) = \alpha(\alpha-1)\cdots(\alpha-k+1)(1+x)^{\alpha-k} \qquad \text{for } x > -1.$$

The nth degree Taylor polynomial of f at 0 is

$$p_n(x) = \sum_{k=0}^{n} \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!} x^k; \tag{14.4}$$

the coefficient of x^k is completely analogous to the combinatorial binomial coefficient,[4] and is therefore read "α choose k" for general α. Three cases deserve further mention:

- $\alpha = -1$. In this case, $f(x) = (1+x)^{-1} = 1/(1+x)$, and the coefficient of x^k is $\big((-1)(-2)(-3)\cdots(-k)\big)/k! = (-1)^k$, so

$$p_n(x) = 1 - x + x^2 - x^3 + \cdots + (-x)^n.$$

[4] That is, if α were a non-negative integer, then the coefficient of x^k would be a combinatorial binomial coefficient.

- $\alpha = 1/2$. In this case, $f(x) = \sqrt{1+x}$, and the coefficient of x^k is
$$\frac{(1/2)(-1/2)(-3/2)(-5/2)\cdots\big((3-2k)/2\big)}{k!} = (-1)^{k-1}\frac{(2k-3)!!}{2^k\,k!},$$
so the Taylor polynomial is
$$p_n(x) = 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} - \frac{5x^4}{128} + \cdots - \frac{(2n-3)!!}{2^n\,n!}(-x)^n.$$
- $\alpha = -1/2$. Here, $f(x) = (1+x)^{-1/2} = 1/\sqrt{1+x}$, and the coefficient of x^k is
$$\frac{(-1/2)(-3/2)(-5/2)\cdots\big((1-2k)/2\big)}{k!} = (-1)^k\frac{(2k-1)!!}{2^k\,k!};$$
the Taylor polynomial is
$$p_n(x) = 1 - \frac{x}{2} + \frac{3x^2}{8} - \frac{5x^3}{16} + \frac{35x^4}{128} - \cdots + \frac{(2n-1)!!}{(2n)!!}(-x)^n.$$

When the exponent α is a non-negative integer (so the power function is a polynomial of degree α), the Taylor polynomials "stabilize" in degree α; otherwise, we get an infinite sequence of distinct polynomials. □

In the previous examples, the Taylor polynomials were found directly from the definition. There are situations in which a direct approach is not feasible, and subterfuge is required. You should not get the impression that a Taylor polynomial *must* be computed from the definition; see Corollary 14.10.

Order of Contact

Let f and g be functions defined on some neighborhood of x_0. Recall that big-O and little-o notation give a measure of how close f and g are. Let ϕ be a function defined on some neighborhood of x_0, and let n be a positive integer. We say $\phi = O(|x-x_0|^n)$ if there exists a constant C and a $\delta > 0$ such that
$$|\phi(x)| \le C|x-x_0|^n \qquad \text{for } |x-x_0| < \delta.$$
We say $\phi = o(|x-x_0|^n)$ if
$$\lim_{x\to x_0}\frac{\phi(x)}{(x-x_0)^n} = 0.$$

Clearly

$$\phi = o(|x - x_0|^n) \implies \phi = O(|x - x_0|^n) \implies \phi = o(|x - x_0|^{n-1}).$$

Neither implication is reversible in general: $|x|^2$ is $O(|x|^2)$ but not $o(|x|^2)$, while $|x|^{3/2}$ is $o(|x|)$ but not $O(|x|^2)$.

Two n-times differentiable functions f and g are *equal to order n at x_0* if $f - g = o(|x - x_0|^n)$, namely if

$$\lim_{x \to x_0} \frac{f(x) - g(x)}{(x - x_0)^n} = 0.$$

Sometimes one says the graphs of f and g have *order-n contact* at x_0 in this situation. Order-0 contact means the graphs cross, while order-1 contact means the graphs are tangent. Higher order of contact is regarded as "higher-order tangency."

We next quantify the claim that the degree n Taylor polynomial of an n-times differentiable function f is the "best" approximation. In words, the next two theorems assert that f and p_{n,x_0} have order-n contact at x_0, and the Taylor polynomial is the *only* polynomial of degree at most n that has this property.

Theorem 14.8. *If f is $\mathscr{C}^n$ near x_0, then $f - p_{n,x_0} = o\big(|x - x_0|^n\big)$.*

Proof. Because the kth derivatives of f and p_n are continuous and agree at x_0 for $k \leq n$, we may apply l'Hôpital's rule n times:

$$\lim_{x \to x_0} \frac{f(x) - p_{n,x_0}(x)}{(x - x_0)^n} = \lim_{x \to x_0} \frac{f^{(n)}(x) - p^{(n)}_{n,x_0}(x)}{n!} = 0,$$

which completes the proof. □

Theorem 14.9. *If p and q are polynomials of degree at most n, and if $p - q$ is $o(x - x_0)^n$ for some x_0, then $p = q$.*

Proof. The hypothesis implies $p - q$ is $o(x - x_0)^k$ for $k = 0, \ldots, n$. Write $(p - q)(x) = b_0 + b_1(x - x_0) + \cdots + b_n(x - x_0)^n$. Taking $k = 0$ implies $b_0 = 0$. Crossing off the first term and taking $k = 1$ implies $b_1 = 0$. Proceeding successively shows that $b_k = 0$ for all $k \leq n$, which means $p = q$. □

A formal proof by induction on n is straightforward, and should be provided if you are bothered by the informality of the argument. For the record, here is the useful consequence, which may be termed "uniqueness of Taylor polynomials."

Corollary 14.10. *If f is n-times continuously differentiable at x_0, and if p is a polynomial having order n contact with f at x_0, then $p = p_{n,x_0}$.*

Corollary 14.11. *If f is real-analytic, then the Taylor polynomial of degree n is obtained by truncating the power series at degree n.*

Proof. By Corollary 11.20,

$$\begin{aligned} f(x) = \sum_{k=0}^{\infty} a_k(x-x_0)^k &= \sum_{k=0}^{n} a_k(x-x_0)^k + \sum_{k=n+1}^{\infty} a_k(x-x_0)^k \\ &= \sum_{k=0}^{n} a_k(x-x_0)^k + O\big(|x-x_0|^{n+1}\big), \end{aligned}$$

so Corollary 14.10 implies the finite sum is the Taylor polynomial. □

In Chapter 3, we claimed there was an easy calculational procedure for expanding a polynomial in powers of $(x-a)$. As an application of Corollary 14.10, we describe this procedure.

Example 14.12 Let $p(x) = \sum_{k=0}^{n} a_k x^k = \sum_{k=0}^{n} b_k(x-a)^k$. By Corollary 14.10, the polynomial on the right is the degree n Taylor polynomial of f at a, whose coefficients are given by $b_k = p^{(k)}(a)/k!$.

For example, suppose we want to write $p(x) = (x+1)^4$ in powers of $(x-1)$. Taking $a = 1$, we calculate

$$\begin{array}{lll} f(x) = (x+1)^4 & f(1) = 2^4 = 16 & b_0 = f(1) = 16 \\ f'(x) = 4(x+1)^3 & f'(1) = 4 \cdot 2^3 = 32 & b_1 = f'(1)/1! = 32 \\ f''(x) = 12(x+1)^2 & f''(1) = 12 \cdot 2^2 = 48 & b_2 = f''(1)/2! = 24 \\ f'''(x) = 24(x+1) & f'''(1) = 24 \cdot 2 = 48 & b_3 = f'''(1)/3! = 8 \\ f^{(4)}(x) = 24 & f^{(4)}(1) = 24 & b_4 = f^{(4)}(1)/4! = 1 \end{array}$$

We immediately read off

$$(x+1)^4 = 16 + 32(x-1) + 24(x-1)^2 + 8(x-1)^3 + (x-1)^4,$$

an identity that may be checked (laboriously) by expansion. □

Corollary 14.10 can be used to calculate Taylor polynomials indirectly.

Example 14.13 Consider the problem of calculating the Taylor polynomials of $f = \mathrm{Tan}^{-1}$ at $x_0 = 0$. The direct approach is a dead end; the first few derivatives are

$$f'(x) = \frac{1}{x^2+1}, \quad f''(x) = -\frac{2x}{(x^2+1)^2}, \quad f'''(x) = \frac{6x^2-2}{(x^2+1)^3}, \ \ldots$$

and there is neither an easily discernible pattern nor a simplification that allows the general derivative to be found. Instead, we reason as follows. By the finite geometric sum formula, if n is a positive integer, then

$$\begin{aligned}\frac{1}{1+t^2} &= 1 - t^2 + t^4 - \cdots + (-1)^n t^{2n} + \frac{(-1)^{n+1}t^{2n+1}}{1+t^2} \\ &= \left(\sum_{k=0}^{n}(-1)^k t^{2k}\right) + \frac{(-1)^{n+1}t^{2n+2}}{1+t^2} \qquad \text{for all } t \in \mathbf{R}.\end{aligned}$$

Integrating this from 0 to x gives

$$\operatorname{Tan}^{-1}x = \int_0^x \frac{1}{1+t^2}dt = \left(\sum_{k=0}^{n}(-1)^k \frac{x^{2k+1}}{2k+1}\right) + \int_0^x \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt.$$

The term in parentheses is a polynomial of degree $(2n+1)$, and the integral is the "error term" to be estimated. Now,

$$\left|\int_0^x \frac{(-1)^{n+1}t^{2n+2}}{1+t^2}\,dt\right| \le \int_0^{|x|}\left|\frac{t^{2n+2}}{1+t^2}\right|dt \le \int_0^{|x|}\left|t^{2n+2}\right|dt = \frac{|x|^{2n+3}}{2n+3},$$

which proves the error term is $o(|x|^{2n+2})$. We have written Tan^{-1} as the sum of a polynomial of degree $(2n+1)$ and an error term that is $o(|x|^{2n+1})$. By Corollary 14.10, the polynomial is the degree-$(2n+1)$ Taylor polynomial of Tan^{-1} at 0. □

This example has a historical coda. Setting $x = 1$ in the preceding discussion gives

$$\frac{\pi}{4} = \operatorname{Tan}^{-1}1 = \sum_{k=0}^{n}(-1)^k \frac{1}{2k+1} + A\left(\frac{1}{2n+3}\right)$$

As n increases, the absolute value of the error term decreases to 0, so

$$\frac{\pi}{4} = \sum_{k=0}^{\infty}(-1)^k \frac{1}{2k+1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots. \tag{14.5}$$

Unfortunately, this remarkable equation does not furnish a good numerical technique for calculating π; the error term is far too large. To guarantee three decimal places, the error must be no larger than $0.5 \times 10^{-3} = 1/2000$. The series above does not achieve this accuracy

until 999 terms have been summed.[5] Worse, each additional decimal place requires summing *ten times as many terms*. There are better numerical techniques for calculating π even from this series, see Exercise 14.13. The general idea is to use trigonometric identities to relate known constants to the value of Tan^{-1} at a number of small absolute value; the $|x|^{2n+3}$ makes the error small for relatively small n.

The Remainder Term

The difference between a function f and one of its Taylor polynomials is called a *remainder term*. Remember that the general problem of this chapter is the numerical evaluation of non-rational functions, and the strategy is to write f as a polynomial plus an error term that is easy to estimate. "Taylor's theorem" allows remainder terms to be estimated systematically, provided enough information about f is known. There are three standard expressions for a remainder, called the integral form, the Lagrange form, and the Cauchy form. Each is useful in certain applications; we will discuss only the integral and Lagrange forms, as they are adequate for our purposes and easy to remember. We also do not attempt to give the weakest differentiability hypotheses, since most of the functions of interest to us are smooth.

Theorem 14.14. *Suppose f is of class $\mathscr{C}^{n+1}$ on some neighborhood $N_\delta(x_0)$, and write $R_{n,x_0}(x) = f(x) - p_{n,x_0}(x)$ for $|x - x_0| < \delta$. Then*

$$R_{n,x_0}(x) = \int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x-t)^n\,dt \qquad \text{(Integral Form)},$$

and there exists z with $|z - x_0| < \delta$ such that

$$R_{n,x_0}(x) = \frac{f^{(n+1)}(z)}{(n+1)!}(x-x_0)^{n+1} \qquad \text{(Lagrange Form)}.$$

Qualitatively, if an $(n+1)$-times differentiable function f is to be approximated as accurately as possible near x_0 by a polynomial of degree at most n, then the best strategy is to use the Taylor polynomial of degree n, and in this case

$$f(x) = p_{n,x_0}(x) + O\big(|x-x_0|^{n+1}\big).$$

Theorem 14.14 gives specific bounds on the constant in the O.

[5] We have proven only that 999 terms suffice to give 3 decimal places, but it is not difficult to show that in this example the error term is no *smaller* than $1/(4n+6)$, so *at least* 499 terms are needed.

Proof. The second fundamental theorem says

$$P(0): \qquad f(x) = f(x_0) + \int_{x_0}^{x} f'(t)\,dt \qquad \text{for } |x - x_0| < \delta,$$

which is precisely the integral form of the remainder for $n = 0$. Assume inductively that

$$P(n): \qquad f(x) = \underbrace{\sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k}_{p_{n,x_0}(x)} + \underbrace{\int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x - t)^n\,dt}_{R_{n,x_0}(x)}$$

Integrate the remainder term by parts, using

$$\begin{aligned} u(t) &= f^{(n+1)}(t) & v(t) &= -\frac{1}{(n+1)!}(x-t)^{n+1} \\ u'(t) &= f^{(n+2)}(t) & v'(t) &= \frac{1}{n!}(x-t)^n\,dt \end{aligned}$$

Since $\int uv' = uv - \int v'u$, the integral becomes

$$\begin{aligned} R_{n,x_0}(x) &= \int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x-t)^n\,dt \\ &= -\frac{f^{(n+1)}(t)}{(n+1)!}(x-t)^{n+1}\Big|_{t=x_0}^{t=x} + \int_{x_0}^{x} \frac{f^{(n+2)}(t)}{(n+1)!}(x-t)^{n+1}\,dt \\ &= \frac{f^{(n+1)}(x_0)}{(n+1)!}(x-x_0)^{n+1} + \int_{x_0}^{x} \frac{f^{(n+2)}(t)}{(n+1)!}(x-t)^{n+1}\,dt. \end{aligned}$$

Adding to $P(n)$ gives $P(n+1)$. By induction, the integral form of the remainder is established:

$$R_{n,x_0}(x) = \int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x-t)^n\,dt \qquad \text{for all } n \geq 0.$$

To obtain the Lagrange form of the remainder, let M and m denote the maximum and minimum of $f^{(n+1)}$ between x_0 and x. Monotonicity of the integral implies

$$\frac{m}{(n+1)!}(x-x_0)^{n+1} \leq R_{n,x_0}(x) \leq \frac{M}{(n+1)!}(x-x_0)^{n+1}.$$

Consequently, there exists c with $m \le c \le M$ and

$$R_{n,x_0}(x) = \frac{c}{(n+1)!}(x - x_0)^{n+1}.$$

By the intermediate value theorem, there exists z between x_0 and x such that $c = f^{(n+1)}(z)$. □

For the basic elementary functions (exp, sin, and cos), Taylor's theorem gives effective bounds on the error term, which in turns says *quantitatively* how good a specific polynomial approximation is.

Example 14.15 Let $f = \exp$ be the natural exponential function. Since $f' = f$, induction on k shows that $|f^{(k)}(x)| = |e^x| \le e$ for all $k \in \mathbf{N}$ and $|x| \le 1$. Example 14.4 and the Lagrange form of the remainder imply that for all $n \in \mathbf{N}$ and $|x| \le 1$,

$$e^x = \sum_{k=0}^{n} \frac{x^k}{k!} + A\left(\frac{e\,|x|^{n+1}}{(n+1)!}\right).$$

We know from Exercise 7.17 of Chapter 7 that $e < 3$. The previous equation immediately implies equation (14.2). See Figure 14.1 for a geometric interpretation of these error bounds. Note that the error of the estimate (the height of the shaded region) increases dramatically away from $x = 0$.

When $|x| \le 1$, the exponential series converges rapidly; for example, $1/10! < 2.76 \times 10^{-7}$, so using terms up to degree 9 to estimate $e^{1/2}$ gives an error smaller than

$$\frac{3 \cdot (1/2)^{10}}{10!} < 8.1 \times 10^{-10}.$$

Squaring the result gives $e = (e^{1/2})^2$ to an accuracy better than 5×10^{-9}, or 8 decimals. □

Example 14.16 For the sine function, there are no terms of even degree, and all derivatives are bounded by 1 in absolute value, so

$$\begin{aligned} |R_{2n+1}(x)| = |R_{2n+2}(x)| &= \left| \int_0^x \frac{\sin^{(2n+3)}(t)}{(2n+2)!}(x-t)^{2n+2}\,dt \right| \\ &\le \left| \int_0^x \frac{(x-t)^{2n+2}}{(2n+2)!}\,dt \right| \quad \le \quad \frac{|x|^{2n+3}}{(2n+3)!}. \end{aligned}$$

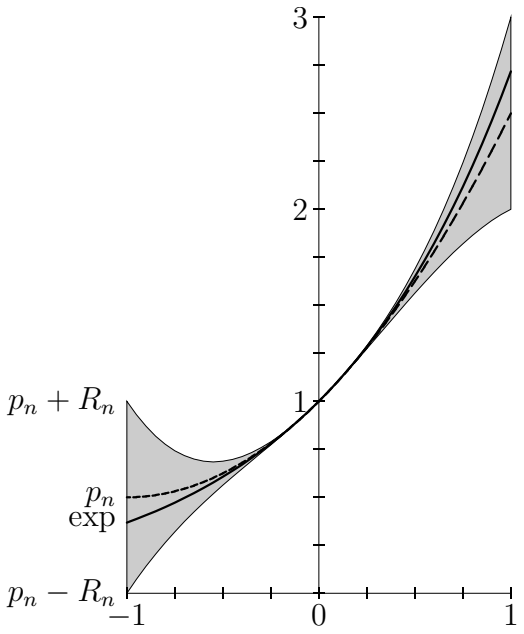


Figure 14.1: Taylor approximation of exp on $[-1, 1]$.

Thus

$$\sin x = \left(\sum_{k=0}^{n} (-1)^k \frac{x^{2k+1}}{(2k+1)!}\right) + A\left(\frac{|x|^{2n+3}}{(2n+3)!}\right). \tag{14.6}$$

Suppose we wish to compute $\sin x$ for x real. Because sin is 2π-periodic, it is enough to have an accurate table for $|x| \le \pi$, and since $\sin(\pi - x) = \sin x$ for all x it is actually sufficient to have an accurate table for $0 < x \le \pi/2$ and to compute π accurately. Recall that $\pi/2 < 1.6$ by a result of Chapter 13. The error bound in Table 14.1, $(1.6)^{2n+3}/(2n+3)!$, is guaranteed by Taylor's theorem.

As a practical matter, to calculate $\sin x$ to 5 decimal places for arbitrary $|x| \le \pi/2$, it is enough to use the degree 9 Taylor polynomial, while the polynomial of degree 13 gives almost 9 decimals. To compute $\sin x$ for x outside this interval, first add or subtract an appropriate multiple of 2π to get a number in $[-\pi, \pi]$, then use the relation $\sin(-x) = -\sin x$ to get a number in $[0, \pi]$, and finally use $\sin(\pi - x) = \sin x$ to get a number in $[0, \pi/2]$. These manipulations depend on having an accurate value for π itself (say 20 decimals), which is easily hard-coded into a calculator chip or computer program. It

Degree $(2n+2)$	$p_{2n+2,0}(x)$	Error Bound
2	x	0.683
4	$x - \frac{x^3}{3!}$	0.0874
6	$x - \frac{x^3}{3!} + \frac{x^5}{5!}$	0.00533
8	$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!}$	0.00019
10	$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!}$	0.00000441

Table 14.1: Approximating $\sin x$ with Taylor polynomials.

should also be noted that for x close to 0, considerably fewer terms are needed to get similar accuracy. For example,

$$\sin(0.1) = (0.1) - \frac{(0.001)}{6} + \text{error}, \qquad |\text{error}| \le \frac{(0.1)^5}{5!} < 10^{-7};$$

the third-degree Taylor polynomial furnishes 6 decimals of $\sin(0.1)$. □

Example 14.17 Let $f : [-1, 1) \to \mathbf{R}$ be defined by $f(x) = \log(1 - x)$. The Taylor polynomials may be found either directly (Exercise 14.5), or by the same trick that was used for Tan^{-1}:

$$f'(x) = -\frac{1}{1-x} = -\left(\sum_{k=0}^{n} x^k\right) - \frac{x^{n+1}}{1-x},$$

while $f(0) = 0$, so

$$f(x) = \int_0^x f'(t)\,dt = -\left(\sum_{k=0}^{n} \frac{x^{k+1}}{k+1}\right) - \int_0^x \frac{t^{n+1}}{1-t}\,dt.$$

The error term is increasingly badly-behaved as $x \to 1$ (as should be expected, since f is unbounded near 1), but there is a simple estimate

$$\left|\int_0^x \frac{t^{n+1}}{1-t}\,dt\right| \le \left(\frac{1}{1-a}\right)\frac{|x|^{n+2}}{n+2} \qquad \text{for } x \le a < 1.$$

To compute $\log 2$, for example, we might first set $a = 1/2$, then compute $\log 1/2 = -\log 2$ by evaluating the series at $x = 1/2$; this utilizes the factor of $(1/2)^{n+1}$ in the error bound. The "obvious" approach, setting $x = -1$, gives the error bound $1/(n+1)$, which decreases very slowly, similarly to the bound in (14.5). However, setting $x = -1$ gives another famous evaluation, the sum of the alternating harmonic series:

$$\log 2 = \sum_{k=0}^{\infty} \frac{(-1)^k}{k+1} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$$

Note carefully that while

$$\log(1-x) = -\sum_{k=0}^{\infty} \frac{x^{k+1}}{k+1} \qquad \text{for } -1 < x < 1,$$

it does not immediately follow that equality holds at $x = -1$, even though the function on the left is continuous and the series on the right is convergent at $x = -1$: The convergence on the right is not uniform near $x = -1$. □

The Binomial Series

Let α be a real number that is *not* a positive integer. As calculated in equation (14.4), the function $f(x) = (1+x)^\alpha$ has Taylor polynomials

$$p_n(x) = 1 + \sum_{k=1}^{n} \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!} x^k.$$

The ratio test implies that the resulting series

$$g(x) = 1 + \sum_{k=1}^{\infty} \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!} x^k =: 1 + \sum_{k=1}^{\infty} a_k\, x^k$$

has radius

$$\lim_{k\to\infty} \left| \frac{a_k}{a_{k+1}} \right| = \lim_{k\to\infty} \left| \frac{k+1}{\alpha - k} \right| = 1,$$

so the domain of g contains the open interval $(-1, 1)$. Aside from technical details similar to what we have done for exp and the circular trig functions, this proves Newton's binomial theorem:

Theorem 14.18. *If α is not a non-negative integer, then*

$$(1+x)^\alpha = 1 + \sum_{k=1}^{\infty} \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!} x^k \qquad \text{for } |x| < 1.$$

The remaining steps are sketched in Exercise 14.14.

Exercises

Exercise 14.1 Let $a \in \mathbf{R}$. Compute the Taylor polynomial of exp, using a as the center of expansion. Conclude that $e^x = e^a e^{x-a}$. ◇

Exercise 14.2 Use Taylor's theorem to estimate the error in approximating $\cos x$ by the value of its degree $2n$ Taylor polynomial. How many terms are needed to estimate $\cos 1$ to 4 decimal places? ◇

Exercise 14.3 Find the Taylor polynomials of cosh. You should be able to guess what they are before calculating anything. ◇

Exercise 14.4 Find the fourth and fifth degree Taylor polynomials of tan directly from the definition. ◇

Exercise 14.5 Let $f(x) = \log(1-x)$ for $|x| < 1$. Find the degree n Taylor polynomial of f *from the definition.* ◇

Exercise 14.6 Let $f(x) = \log(1+x^2)$ for $|x| < 1$. Find the general Taylor polynomial of f. You are cautioned that the direct method is not the easiest approach. ◇

Exercise 14.7 Compare the Taylor polynomials p_n found in Exercise 14.5 with the polynomials q_n found in Exercise 14.6: $q_n(x) = p_n(-x^2)$. State and prove a theorem relating the Taylor polynomials of f and $g(x) = f(cx^2)$, with $c \in \mathbf{R}$. ◇

Exercise 14.8 Find the Taylor series of the following functions:

(a) $f_1(x) = e^{-x^2}$ (b) $f_2(x) = \sin(x^2)$ (c) $f_3(x) = \cos(x^2)$.

Suggestion: Use the previous exercise. ◇

Exercise 14.9 Use series products to verify that

$$\frac{1}{1-x^2} = \frac{1}{1+x} \cdot \frac{1}{1-x}$$

for $|x| < 1$. ◇

Exercise 14.10 Find the Taylor series of

$$\tanh x = \int_0^x \frac{1}{1-t^2}\,dt.$$

For which x does the series converge? ◇

Exercise 14.11 Let

$$f(x) = \int_0^x e^{-t^2}\,dt.$$

(a) Find the Taylor series of f. For which x does the series converge?

(b) Using only the arithmetic functions on a calculator, evaluate $f(1)$ to 4 decimals.
Hint: How many terms of the series are needed?

(c) Prove that $\lim(f, +\infty)$ exists. This question is unrelated to the material on Taylor series. Explain carefully why series are no help in computing an improper integral.

A function closely related to f arises in probability. ◇

Exercise 14.12 Use Taylor's theorem to bound the nth remainder term for exp. When $x = 1$, is this estimate better or worse than the estimate obtained with series in Chapter 12? ◇

Exercise 14.13 Use $\tan \frac{\pi}{6} = \frac{1}{\sqrt{3}}$ to prove that

$$\frac{\pi\sqrt{3}}{6} = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)3^k}$$

How many terms are needed to guarantee 3 decimals of accuracy? ◇

Exercise 14.14 This exercise outlines the proof of Theorem 14.18. Suppose α is not a non-negative integer, and set $f(x) = (1+x)^\alpha$ and

$$g(x) = 1 + \sum_{k=1}^{\infty} \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!} x^k \qquad \text{for } |x| < 1.$$

(a) Use series manipulations to prove that $(1+x)g'(x) = \alpha g(x)$ on $(-1, 1)$.

(b) Use part (a) to prove that g/f is constant on $(-1, 1)$.

It may help to review the uniqueness proof for exp in Chapter 12. ◇

Exercise 14.15 Prove that the property of being "equal to order r at x_0" is an equivalence relation on the space of smooth functions defined on some neighborhood of x_0. An equivalence class is called an "r-jet" at x_0. ◇

Translation Exercises

How well have you assimilated the language of mathematics? For example, when you hear, "The total contribution of the terms f_n becomes negligible over the whole interval", you should think:

> Let $I \subset \mathbf{R}$ be an interval, and let (f_n) be a sequence of functions on I. For every $\varepsilon > 0$, there exists a positive integer N such that $\sum_{n=N+1}^{\infty} |f_n(x)| < \varepsilon$ for all x in I.

Here is your chance to match wits with Goethe:

Exercise 14.16 Translate each of the following informal phrases or sentences into a precise, quantified statement.

- If (a_n) and (b_n) are sequences of positive terms, and if a_n/b_n approaches $1/2$ as $n \to \infty$, then $0 < a_n < b_n$ for n sufficiently large.
- The function value $f(x)$ is close to $f(x_0)$ whenever x is close to x_0.
- The function $f(x) = x^{-(1+1/x)}$ is asymptotic to $1/x$ as $x \to \infty$.
- For $x \simeq 0$, $1 - \cos x$ is much smaller than x.
- Adding up sufficiently many terms a_n, the partial sums may be brought as close to the sum as desired.
- Letting the widths of the inscribed rectangles approach zero, the area under the graph of $f : [a, b] \to \mathbf{R}$ may be approximated arbitrarily closely.
- The terms a_n may be arranged arbitrarily without changing the sum of the series.
- The terms a_n eventually decrease to zero.

If appropriate, identify the property or theorem. ◇

Chapter 15

Elementary Functions

Recall that a function f is *real-analytic* at x_0 if there exists a power series with positive radius of convergence that is equal to f on some neighborhood of x_0. For every a in the interval of convergence, the power series can be expended at a, and the radius of convergence of the new series is positive. Therefore, a function that is analytic at a point is analytic in an open interval.

Our library of analytic functions includes

- Algebraic functions (functions defined implicitly by a polynomial in two variables), such as $f(x) = \sqrt[3]{1+x^2}/\sqrt{3-x}$, which satisfies

$$F(x,y) = (1+x^2)^2 - (3-x)^3 y^6 = 0.$$

  One-variable polynomials and rational functions fall into this category.

- Exponential and logarithmic functions, and hyperbolic trig functions.

- Circular trig functions.

A function that can be obtained from these families by finitely many arithmetic operations and function compositions is said to be *elementary*. "Typical" elementary functions include $f_1(x) = \log\bigl(x+\sqrt{x^2-1}\bigr)$,

$$f_2(x) = x^x = e^{x\log x}, \qquad f_3(x) = e^{e^{\sin[1+\sqrt{\log(4+x^2)}]}}.$$

15.1 A Short Course in Complex Analysis

The circular trig functions bear a strong formal similarity to the hyperbolic trig functions, which are clearly related to the exponential function. To explain the magic underlying this similarity, it is necessary to work over the complex numbers, specifically to consider complex power series and complex differentiation. To cover this material in full detail is beyond the scope of this book, but the simpler aspects are almost completely analogous to definitions and theorems we have already introduced.

Complex Arithmetic

Let a and b be real numbers. The *norm* of $\alpha := a + bi \in \mathbf{C}$ is

$$|\alpha| := \sqrt{\alpha\bar{\alpha}} = \sqrt{a^2 + b^2},$$

namely the distance from 0 to α. The norm function on $\mathbf{C}$ has properties very similar to the absolute value function on $\mathbf{R}$; aside from the obvious fact that the complex norm extends the real absolute value, the norm is multiplicative and satisfies triangle inequalities:

Theorem 15.1. *Let α and β be complex numbers. Then $|\alpha\beta| = |\alpha|\,|\beta|$, and*

$$\Big||\alpha| - |\beta|\Big| \leq |\alpha + \beta| \leq |\alpha| + |\beta|. \tag{15.1}$$

These inequalities are the *reverse triangle* and *triangle* inequalities.

Proof. Write $\alpha = a + bi$ and $\beta = x + yi$, with a, b, x, and y real. By definition,

$$\alpha\beta = (a + bi)(x + yi) = (ax - by) + (ay + bx)i,$$

so a direct calculation gives

$$\begin{aligned} |\alpha\beta|^2 &= (ax - by)^2 + (ay + bx)^2 \\ &= (ax)^2 - 2(axby) + (by)^2 + (ay)^2 + 2(aybx) + (bx)^2 \\ &= (a^2 + b^2)(x^2 + y^2) \quad = \quad |\alpha|^2\,|\beta|^2. \end{aligned}$$

The triangle inequality may also be established with a brute-force calculation, but it is more pleasant to use complex formalism. For all $\alpha \in \mathbf{C}$, $|\alpha + \bar{\alpha}| = |2\,\mathrm{Re}\;\alpha| \leq 2|\alpha|$. In particular,

$$|\alpha\bar{\beta} + \beta\bar{\alpha}| = |\alpha\bar{\beta} + \overline{\alpha\bar{\beta}}| \leq 2|\alpha\bar{\beta}| = 2|\alpha|\,|\bar{\beta}| = 2|\alpha|\,|\beta|$$

for all α, $\beta \in \mathbf{C}$. Thus

$$\begin{aligned} |\alpha+\beta|^2 = (\alpha+\beta)\overline{(\alpha+\beta)} &= |\alpha|^2 + \alpha\bar{\beta} + \beta\bar{\alpha} + |\beta|^2 \\ &\leq |\alpha|^2 + 2|\alpha|\,|\beta| + |\beta|^2 = (|\alpha| + |\beta|)^2. \end{aligned}$$

Taking square roots proves the triangle inequality for complex numbers. The reverse triangle inequality is established by the same trick as was used to prove the real version, Theorem 2.25. □

Armed with these basic tools, we could, in principle, return to Chapters 4, 8 and 11, and check that the definitions and properties of sequences, limits, continuity, differentiability, power series, and radius of convergence can be made in exactly the same manner for complex-valued functions of a complex variable, provided we interpret absolute values as norm of complex numbers. The domain of convergence of a power series centered at $a \in \mathbf{C}$ is a *bona fide* disk in the complex plane, explaining the term "radius" of convergence.

Integration is conspicuously absent from the above list of items that generalize immediately. The basic reason is that integration relies heavily on order properties of $\mathbf{R}$; partitions of intervals are defined using the ordering, and upper and lower sums have no meaning for complex-valued functions, since sup and inf are concepts requiring ordering. There is a different kind of integration—"contour integration"—that does play a central role in complex analysis, but its definition and significance are beyond the scope of this book.

Exp and the Trig Functions

For the remainder of this section, we restrict attention to complex power series with *infinite* radius of convergence. A function $f : \mathbf{C} \to \mathbf{C}$ associated to such a power series is said to be *entire*; the examples we have seen so far are polynomials, exp and the hyperbolic trig functions sinh and cosh, the circular trig functions sin and cos, and functions built from these by adding, multiplying, and composing functions finitely many times. The complex power series of the "basic" elementary functions are identical to the real power series found earlier; the only difference is that the "variable" is allowed to be complex (and is traditionally denoted z instead of x).

Compare the power series of cos and cosh:

$$\cos z = \sum_{k=0}^{\infty} (-1)^k \frac{z^{2k}}{(2k)!} = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \cdots$$
$$\cosh z = \sum_{k=0}^{\infty} \frac{z^{2k}}{(2k)!} = 1 + \frac{z^2}{2!} + \frac{z^4}{4!} + \frac{z^6}{6!} + \cdots$$

Both series have radius $+\infty$, just as in the real case, so each represents an entire function. A moment's thought reveals that $\cos z = \cosh(iz)$, since $(iz)^{2k} = (-1)^k z^{2k}$. The simple expedient of multiplying the variable by i converts cosh to cos! Analogously,

$$\sin z = \sum_{k=0}^{\infty} (-1)^k \frac{z^{2k+1}}{(2k+1)!} = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots$$
$$\sinh z = \sum_{k=0}^{\infty} \frac{z^{2k+1}}{(2k+1)!} = z + \frac{z^3}{3!} + \frac{z^5}{5!} + \frac{z^7}{7!} + \cdots$$

The relation between them is $i \sin z = \sinh(iz)$ by similar reasoning.

Recall the the circular and hyperbolic trig functions were characterized as solutions of certain initial value problems. The reason for the similarity of the differential equations that characterize these functions should not be difficult to see. Suppose we consider the function $f(z) = \sinh(kz)$ for some $k \in \mathbf{C}$. Differentiating gives $f'(z) = k \cosh(kz)$, so $f'(0) = k$, and then $f''(z) = k^2 \sinh(kz)$; in summary,

$$f'' - k^2 f = 0, \qquad f(0) = 0, \quad f'(0) = k.$$

Now, suppose we take $k = i$; the preceding equation becomes $f'' + f = 0$, $f(0) = 0$, $f'(0) = i$. But the characterization of sin and cos as solutions of the equation $f'' + f = 0$ (Corollary 13.3, whose proof generalizes immediately to the complex domain) shows that $f(z) = i \sin z$, and again we find that $i \sin z = \sinh(iz)$. Analogous remarks are true for cos and cosh.

If we assemble the conclusions of the preceding paragraphs, we obtain an identity usually called (when z is real) *de Moivre's theorem*:

Theorem 15.2. $e^{iz} = \cos z + i \sin z$ *for all* $z \in \mathbf{C}$.

The special case $z = \pi$ (usually written $e^{i\pi} + 1 = 0$) is called *Euler's formula*. It is too much to say this equation has mystical significance,

but it strikingly relates "five of the most important constants in mathematics."

The proof of Theorem 15.2 is one line:

$$e^{iz} = \cosh(iz) + \sinh(iz) = \cos z + i \sin z.$$

The same conclusion results by direct inspection of the respective power series. This formula, from the point of view of real variables, is nothing short of incredible. What does geometric growth have to do with measuring circles? The answer is hidden in the nature of complex multiplication, as will gradually become apparent. To give one link, observe that if x and y are real, then

$$\begin{aligned} e^{ix}e^{iy} &= (\cos x + i \sin x)(\cos y + i \sin y) \\ &= (\cos x \cos y - \sin x \sin y) + i(\sin x \cos y + \sin y \cos x) \\ &= \cos(x+y) + i \sin(x+y) \qquad \text{by equation (13.9)} \\ &= e^{i(x+y)}. \end{aligned}$$

The addition formulas for sin and cos are nothing but the addition rule for exponentials, extended to the complex numbers! A closely related formula is worth emphasizing.

Corollary 15.3. *If x and y are real, then $e^{x+iy} = e^x \cos y + ie^x \sin y$.*

The Geometry of Complex Multiplication

The set of *unit complex numbers*, $S^1 := \{z \in \mathbf{C} \mid z\bar{z} = 1\}$, is geometrically a circle of radius 1 centered at 0.

Proposition 15.4. *Every unit complex number is of the form e^{it} for a unique $t \in (-\pi, \pi]$.*

Proof. The function $\gamma : \mathbf{R} \to \mathbf{C}$ defined by $\gamma(t) = e^{it}$ is 2π-periodic, and has no smaller positive period. Since $|e^{it}| = \sqrt{\cos^2 t + \sin^2 t} = 1$, the restriction of γ is an injective mapping $(-\pi, \pi] \to S^1$.

To prove that $\gamma|_{(-\pi,\pi]}$ is surjective, recall that cos maps $[0, \pi]$ bijectively to $[-1, 1]$. If $\alpha = a + bi \in S^1$ with $b \geq 0$, then there is a unique $t \in [0, \pi]$ with $a = \cos t$. For this t,

$$b = \sqrt{1 - a^2} = \sqrt{1 - \cos^2 t} = \sin t,$$

so $\alpha = e^{it}$. If instead $b < 0$, then $\bar{\alpha} = e^{it}$ for a unique $t \in (0, \pi)$, so

$$\alpha = \overline{e^{it}} = \cos t - i \sin t = \cos(-t) + i \sin(-t) = e^{-it}.$$

That is, $\alpha = e^{it}$ for a unique $t \in (-\pi, 0)$. Note that by periodicity, γ maps *every* half-open interval of length 2π bijectively to the circle. □

Every non-zero complex number z is a product of a positive real number and a complex number of unit norm: $z = |z| \cdot (z/|z|)$. By Proposition 15.4, we may write

$$z = e^{\rho} \cdot e^{it} = e^{\rho + it} \qquad \text{for } \textit{unique } \rho \in \mathbf{R} \text{ and } t \in (-\pi, \pi]. \tag{15.2}$$

This representation is called the *polar form* of z; the number t is called the *argument* of z, denoted $\arg z$, and is interpreted as the angle between the positive real axis and the ray from 0 to z.

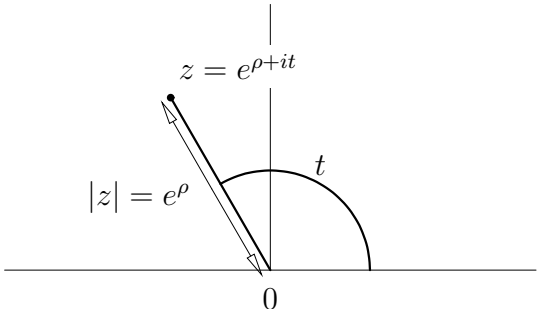


Figure 15.1: The polar form of a non-zero complex number.

Equation (15.2) immediately yields the geometric description of complex multiplication in Figure 2.12 (page 81): If $z_j = e^{\rho_j + it_j}$ for $j = 1, 2$, then

$$z_1 z_2 = (e^{\rho_1} \cdot e^{it_1})(e^{\rho_2} \cdot e^{it_2}) = e^{\rho_1} e^{\rho_2} \cdot e^{i(t_1 + t_2)}.$$

In words, we multiply two complex numbers by multiplying their norms and adding their arguments. If $\alpha = e^{\rho + it}$, then the mapping $z \mapsto \alpha z$ rotates the plane through t radians and scales by a factor of e^{ρ}. In particular, multiplication by $i = e^{i\pi/2}$ rotates the plane a quarter turn counterclockwise.

De Moivre's formula has a beautiful geometric interpretation in terms of complex multiplication. Recall that

$$e^x = \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n \qquad \text{for } x \in \mathbf{R}.$$

Suppose we try to interpret this for $x = i\theta$ pure imaginary. The complex number $1 + (i\theta/n)$ defines a right triangle, on the left in Figure 15.2, whose angle at the origin is very nearly θ/n. (The case $\theta = \pi$ is shown.) Raising this number to the nth power corresponds geometrically to iteratively scaling and rotating this triangle, as in the right half of Figure 15.2. The opposite vertex of the last triangle is very nearly on the ray making an angle of θ with the positive horizontal axis. As $n \to \infty$ it is geometrically apparent that the terminal point of the sequence approaches $\cos\theta + i\sin\theta$, which is therefore equal to $e^{i\theta}$.

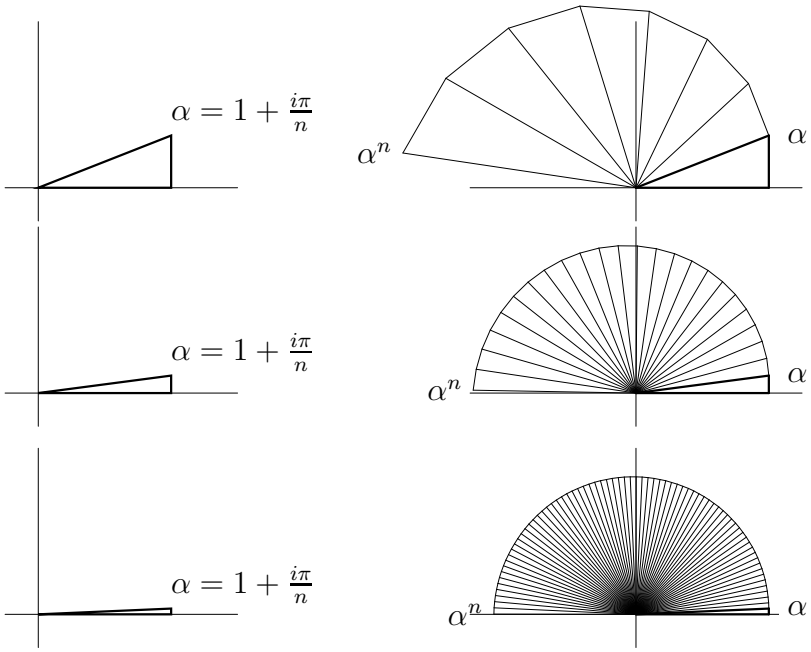


Figure 15.2: The geometric interpretation of de Moivre's formula.

To make this geometric argument precise, write $\alpha = 1 + (i\theta/n) = |\alpha| \cdot e^{it}$ and note that

$$|\alpha| = \sqrt{1 + (\theta/n)^2} = 1 + O(1/n^2), \quad t = \mathrm{Tan}^{-1}(\theta/n) = (\theta/n) + O(1/n^3).$$

The estimates are immediate from the appropriate Taylor polynomials. Taking the nth power, we find that

$$|\alpha^n| = \big(1 + O(1/n^2)\big)^n, \qquad \arg(\alpha^n) = nt = \theta + O(1/n^2).$$

As $n \to +\infty$, we have $|\alpha^n| \to 1$ and $\arg(\alpha^n) \to \theta$, so $\alpha^n \to e^{i\theta}$.

The Complex Logarithm

Recall that $\mathbf{C}^\times$ denotes the set of non-zero complex numbers, which is an Abelian group under complex multiplication. Every w in $\mathbf{C}^\times$ is *uniquely* written as $e^z = e^{\rho+it}$ for ρ real and $t \in (-\pi, \pi]$. The *principle branch of the logarithm* is the mapping $\operatorname{Log} : \mathbf{C}^\times \to \mathbf{C}$ defined by

$$\operatorname{Log} w = z = \rho + it, \quad -\pi < t \leq \pi. \tag{15.3}$$

The image of Log is a horizontal strip of height 2π, the *fundamental strip*, see Figure 15.3. Points on the negative real axis in $\mathbf{C}^\times$ correspond to points with $t = \pi$ in the figure.

If $w \in \mathbf{C}^\times$, then $\exp(\operatorname{Log} w) = w$. By contrast, if $z \in \mathbf{C}$, then in general $\operatorname{Log}(\exp z) = z$ is *false*; the left-hand side has imaginary part in $(-\pi, \pi]$ for all z, so if z does not lie in the fundamental strip, the two sides cannot be equal. Instead, for each z there is a unique integer k such that $z - 2\pi ik$ has imaginary part in $(-\pi, \pi]$, and $\operatorname{Log}(\exp z) = z - 2\pi ik$ for this k.

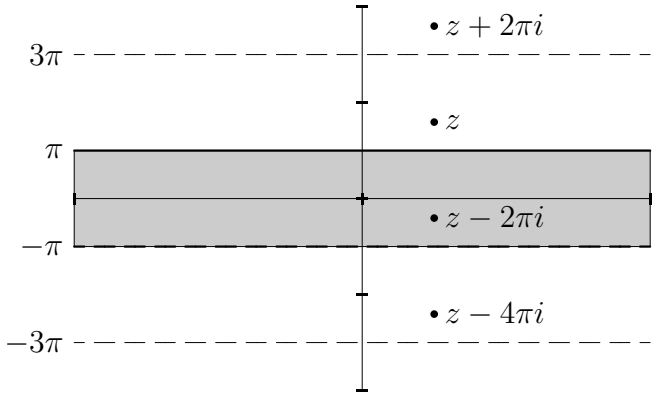


Figure 15.3: The complex logarithm as a mapping.

The complex exponential function is periodic: $e^{z+2\pi i} = e^z$ for all z in $\mathbf{C}$. Thus exp cannot have an inverse function; the best we can hope for is to restrict exp to a set on which it is bijective. The fundamental strip is just such a set. As indicated in Figure 15.3, there are infinitely many "period strips", each corresponding to a branch of logarithm. The points shown all map to the same w under exp, and a branch of Log selects one such point as a function of w.

There is a more subtle issue with invertibility of exp. Consider the sequence (z_n) defined by $z_n = (-\pi + \frac{1}{n})i$. Each term lies in the

fundamental strip, but the limit, $-\pi i$, does not. If we apply exp to this sequence and then Log, we find that

$$\operatorname{Log}(\exp z_n) = z_n, \qquad \text{but} \quad \operatorname{Log}\bigl(\exp(-\pi i)\bigr) = \pi i;$$

the principle branch of the logarithm is *discontinuous*. It is easy to see geometrically that the trouble occurs precisely along the negative real axis in the w plane, which is the image of the line $\operatorname{Im} z = \pi$, the top edge of the fundamental strip. As the point w moves "continuously" across the negative real axis, the point $z = \operatorname{Log} w$ moves "discontinuously" between the top and bottom edges of the fundamental strip.

This discussion is a little informal, since we have not precisely defined the concept of "continuity" for functions of a complex variable. However, it should be plausible that *there does not exist* a continuous branch of logarithm on $\mathbf{C}^\times$. This basic fact is at the origin of varied and subtle phenomena throughout mathematics and physics. Exercise 15.1 gives a whimsical (if mathematically apt) example.

Complex Polynomials

In Chapter 5 we saw that every positive real number has a unique nth root for each positive integer n, and that every polynomial of odd degree has a real root. In this section, we will see how beautifully simple the analogous questions become over the field of complex numbers.

Roots of Complex Numbers

Let w be complex, and let n be a positive integer. An nth root of w is a complex number z satisfying $z^n = w$. There are at most n solutions z of the polynomial equation $z^n - w = 0$, so if we are able to find n distinct nth roots, we have a complete list.

If $z^n - w = 0$ is written out using the real and imaginary parts of z and w, the result is too complicated to yield any obvious insight. However, writing $z = |z|e^{it}$ and $w = |w|e^{i\theta}$ immediately yields n distinct roots when $w \neq 0$. Two numbers in polar form are equal iff they have the same norm and their arguments differ by an integer multiple of $2\pi i$. Consequently, $z^n = w$ iff

$$|z| = |w|^{1/n}, \qquad t = \tfrac{1}{n}(\theta + 2\pi i k), \quad k = 0,\ \ldots,\ n-1.$$

The first is an equation of positive real numbers that we know has a unique solution. The second is an explicit way of listing solutions of $nt - \theta = 2\pi i k$ such that the numbers e^{it} are distinct.

Even the case $w = 1$ is interesting. The numbers $e^{2\pi ik/n}$, $0 \leq k < n$, are called *nth roots of unity*, since $(e^{2\pi ik/n})^n = e^{2\pi ik} = 1$. It is not uncommon to write $\zeta_n := e^{2\pi i/n}$, so that the nth roots of unity are powers of ζ_n. Each root of unity is a unit complex number (i.e., has norm 1), and the angle between consecutive roots is $1/n$th of a turn. These points therefore lie at the vertices of a regular n-gon inscribed in the unit circle:

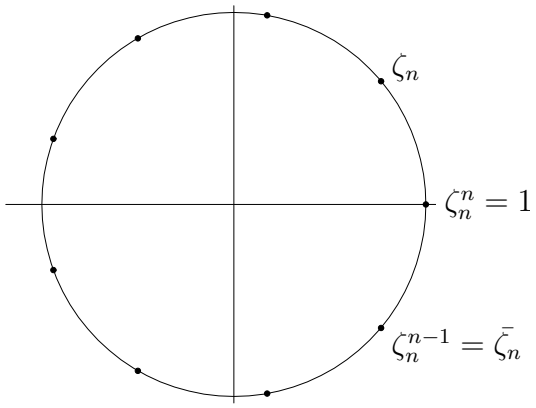


Figure 15.4: The nth roots of unity.

For a general non-zero w, if z is a particular nth root, then the other roots are $z\zeta_n^k$ for $0 \leq k < n$; in particular, the nth roots of w also lie at the vertices of a regular n-gon, inscribed in a circle of radius $|w|^{1/n}$. The n roots of 0 are all 0, but in a meaningful sense there are still n of them!

The Fundamental Theorem of Algebra

The polynomial $p(z) = z^n - w$ factors completely over $\mathbf{C}$: If $z_1, \ldots, z_n$ are the nth roots of w, then

$$z^n - w = \prod_{k=1}^{n} (z - z_k).$$

This property is a special case of the *fundamental theorem of algebra*:

Theorem 15.5. *Let $p : \mathbf{C} \to \mathbf{C}$ be a non-constant polynomial function with complex coefficients. Then there exists a $z_0 \in \mathbf{C}$ with $p(z_0) = 0$.*

Informally, every complex polynomial has a root. Recall that by Corollary 3.7, a root of a polynomial corresponds to a linear factor: $p(z_0) = 0$ iff $(z - z_0)$ divides p. Repeated application of Theorem 15.5 implies that every non-constant polynomial factors completely into linear factors.

Unfortunately, a complete proof requires a technique we have not fully developed, but the sketch presented here should give an idea of what is involved.

Proof. Assuming as usual that the top coefficient a_n is non-zero, write

$$p(z) = \sum_{k=0}^{n} a_k z^k = a_n z^n \big(1 + O(1/z)\big).$$

Taking absolute values, $|p(z)| = |a_n| \cdot |z|^n \big(1 + O(1/|z|)\big)$ near $|z| = +\infty$: A polynomial goes to $+\infty$ in absolute value as $|z| \to +\infty$.

Consider $a_0 = p(0)$; by the previous paragraph, there exists a real number $R > 0$ such that $|p(z)| > |a_0|$ for $|z| > R$. Since $|p| : \mathbf{C} \to \mathbf{R}$ is continuous, the restriction to the disk $D_R := \{z \in \mathbf{C} : |z| \leq R\}$ has an absolute minimum at some $z_0 \in D_R$ by a generalization of the extreme value theorem.[1] The point z_0 is an absolute minimum for $|p|$ on all of $\mathbf{C}$, since by choice of R we have $|p(z)| > |p(0)| \geq |p(z_0)|$ for $|z| > R$. If we can prove that $|p(z_0)| = 0$, we are done.

Expand p in powers of $(z - z_0) = re^{it}$. Letting b_ℓ be the coefficient of the lowest-degree non-constant term, we have

$$p(z) = \sum_{k=0}^{n} b_k (z - z_0)^k = b_0 + b_\ell (z - z_0)^\ell \big(1 + O(z - z_0)\big).$$

Now, if $b_0 \neq 0$, then

$$|p(z_0 + re^{it})| = |b_0| \cdot \left|1 + \frac{b_\ell}{b_0} r^\ell e^{i\ell t} \big(1 + O(r)\big)\right|.$$

Choose $r \ll 1$ such that $r^\ell \big(1 + O(r)\big) > 0$, then choose t so that $(b_\ell/b_0)e^{i\ell t}$ is real and negative. (Writing b_ℓ/b_0 in polar form shows this is possible.) The previous equation implies $|p(z_0 + re^{it})| < |b_0| = |p(z_0)|$, contrary to the fact that z_0 was an absolute minimum of $|p|$. It must be that $p(z_0) = 0$. □

[1] This is the only step we fail to justify substantially. The crucial properties of the disk are that is is a bounded subset of the plane that contains all its limit points.

The fundamental theorem of algebra has interesting consequences for polynomials with real coefficients.

Corollary 15.6. *Let $p : \mathbf{C} \to \mathbf{C}$ be a polynomial with real coefficients. Then $z_0 \in \mathbf{C}$ is a root iff $\bar{z}_0$ is. Further, p factors over the reals into linear and irreducible quadratic terms.*

Proof. Recall that a complex number α is real iff $\bar{\alpha} = \alpha$ and that conjugation commutes with addition and multiplication in the sense of (2.16). Let $p(z) = \sum_k a_k z^k$ with a_k real for all k. Then

$$p(\bar{z}) = \sum_k a_k \bar{z}^k = \sum_k a_k \overline{z^k} = \sum_k \overline{a_k z^k} = \overline{p(z)}.$$

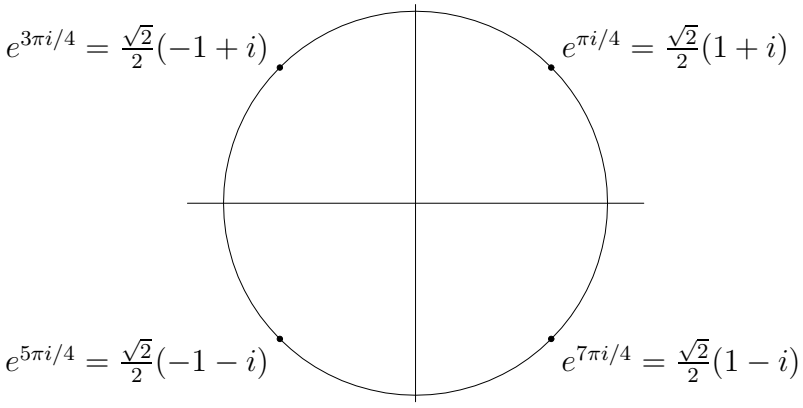
In particular, $p(z_0) = 0$, then $p(\bar{z}_0) = \overline{p(z_0)} = 0$. Since conjugating twice is the identity map, the first claim is proven.

For the second, use the fundamental theorem to factor p into linear terms (with complex coefficients), then group the terms corresponding to conjugate roots. For all $\alpha \in \mathbf{C}$,

$$(z - \alpha)(z - \bar{\alpha}) = z^2 - 2 \operatorname{Re} \alpha + |\alpha|^2$$

has real coefficients, so p is written as a product of real and irreducible quadratic factors with real coefficients. □

Example 15.7 Let $p(z) = z^4 + 1$. The roots of p are the fourth roots of -1, which are 8th roots of 1 that are not square or fourth roots of 1:



Multiplying the corresponding terms in conjugate pairs gives

$$z^4 + 1 = (z^2 + \sqrt{2}z + 1)(z^2 - \sqrt{2}z + 1).$$

In hindsight, we might have written $z^4 + 1 = (z^4 + 2z^2 + 1) - 2z^2$ and used the difference of squares formula to obtain the same result. Note carefully that this polynomial has no real roots, but still factors over the reals. □

15.2 Elementary Antidifferentiation

One of the traditional topics of calculus is the exact evaluation of definite integrals in symbolic terms. This section presents several techniques for calculating elementary antiderivatives, and states (without proof) some criteria under which an elementary function fails to have an elementary antiderivative.

Symbolic calculation of antiderivatives is fundamentally different from symbolic calculation of derivatives. The product, quotient, chain, and implicit differentiation rules imply that the derivative of an elementary function is also elementary, and together provide an effective symbolic calculational tool for differentiating. By contrast, *there does not exist* an algorithm for symbolically antidifferentiating; there are no general formulas analogous to the product, quotient, and chain rules for antiderivatives. Of course, if f is continuous, then

$$F(x) = \int_a^x f(t)\,dt$$

is an antiderivative of f, so every elementary function is, in a sense, trivially antidifferentiable. However, it is a fact of life that an elementary function generally *does not have* an elementary antiderivative, see Theorem 15.10 below, for example.

As computer algebra systems become universal and inexpensive, manual ability with integration techniques will probably disappear gradually, just as modern students have no need for the ability to extract square roots or do long division on paper, thanks to electronic calculators. However, the philosophy of this book is to provide a look at the internal workings of calculus, so at least you should be aware of how computer integration packages work. It is a fringe benefit that some formulas can be used to evaluate to curious improper integrals or infinite sums.

Notation

We shall write

$$F(x) = \int^x f \quad \text{or} \quad F(x) = \int^x f(t)\,dt$$

to indicate that F is the general antiderivative of f. This notation merges well with the notation for "definite" integrals; the constant of integration is absorbed in the missing lower limit, and all occurrences of x in a single equation mean the same thing. As usual, the dummy variable t may be replaced by any non-occurring symbol. One must exercise caution; the expression $\int^x f$ denotes not one function, but an *infinite family* of functions, any two of which differ by an additive constant. The particular antiderivative that vanishes at a is

$$F(x) = \int_a^x f(t)\,dt.$$

This observation is quite useful in practice.

Substitution

The method of substitution arises from the chain rule. Guided by Theorem 10.5, we attempt to find a "change of variable" that simplifies the integrand to the point where it can be antidifferentiated by inspection. Using our notation for antiderivatives, the conclusion of Theorem 10.5 is written

$$\int^x f\big(U(t)\big)U'(t)\,dt = \int^{U(x)} f(u)\,du, \tag{15.4}$$

in which we have used the substitution $u = U(t)$. Without dummy variables, the preceding equation becomes

$$\int^x f(U)\,U' = \int^{U(x)} f.$$

In Leibniz notation, the dummy variables convert satisfyingly because of the custom of writing $u = u(t)$. Formally, we have $du = \frac{du}{dt}\,dt = u'(t)\,dt$, and equation (15.4) arises by (more-or-less literal) symbolic substitution.

Several examples follow; throughout, n is a non-negative integer and $k \neq 0$, $\alpha \neq -1$ are real. The integrals are assumed to be taken

over intervals on which the integrand is defined. For brevity, we give the substitution and its differential, followed by the formula to which it applies. Each "answer" contains an explicit additive constant.

- $u = kt$, $du = k\,dt$.

$$\int^x e^{kt}\,dt = \frac{1}{k}\int^{kx} e^u\,du = \frac{1}{k}e^{kx} + C$$

The same trick handles $\displaystyle\int^x f(kt)\,dt$ if f is antidifferentiable.

- $u = 1 + e^t$, $du = e^t\,dt$.

$$\int^x e^t\sqrt{1+e^t}\,dt = \int^{1+e^x} u^{1/2}\,du = \left.\frac{u^{3/2}}{3/2}\right|^{1+e^x} = \frac{2}{3}(1+e^x)^{3/2} + C$$

- $u = k + t^2$, $du = 2t\,dt$.

$$\int^x (k+t^2)^\alpha t\,dt = \int^{k+x^2} u^\alpha\,\frac{du}{2} = \left.\frac{1}{2}\frac{u^{\alpha+1}}{\alpha+1}\right|^{k+x^2} = \frac{(k+x^2)^{\alpha+1}}{2(\alpha+1)} + C$$

- $u = \cos t$, $du = -\sin t\,dt$.

$$\int^x \cos^n t\,\sin t\,dt = -\int^{\cos x} u^n\,du = \left.-\frac{u^{n+1}}{n+1}\right|^{\cos x} = -\frac{\cos^{n+1} x}{n+1} + C$$

- $u = \cos t$, $du = -\sin t\,dt$.

$$\int^x \frac{\sin t}{\cos t}\,dt = -\int^{\cos x}\frac{du}{u} = \left.-\log|u|\right|^{\cos x} = -\log|\cos x| + C$$

or $\displaystyle\int^x \tan t\,dt = \log|\sec x| + C$.

- The relationship between the complex exponential and the trig functions can be used to evaluate a couple of integrals that otherwise require integration by parts. Let a and b be real, and consider

$$\int^x e^{at}\cos bt\,dt + i\int^x e^{at}\sin bt\,dt = \int^x e^{(a+bi)t}\,dt.$$

The antiderivative formula for exp holds even for complex exponents, so the integral on the right is

$$\frac{1}{a+bi}e^{(a+bi)x} = \frac{a-bi}{a^2+b^2}\,e^{ax}(\cos bx + i\sin bx).$$

Equating real and imaginary parts, we obtain the two useful formulas

$$\int^x e^{at}\cos bt\,dt = \frac{e^{ax}}{a^2+b^2}(a\cos bx + b\sin bx) + C$$

$$\int^x e^{at}\sin bt\,dt = \frac{e^{ax}}{a^2+b^2}(a\sin bx - b\cos bx) + C$$

These functions arise in the study of damped harmonic oscillators.

The method of substitution is something of an art, and only works on a limited set of integrands; there is a big difference between the similar-looking integrals

$$(*) \qquad \int^x e^{-t^2}\,dt \qquad \text{and} \qquad \int^x te^{-t^2}\,dt.$$

The *general* course of action is to let u be whatever is inside a radical or transcendental function, particularly if du is visibly present in the remainder of the integrand. Sometimes an "obvious" substitution leads to an integral requiring a further substitution; in such a case, the same transformation can always be accomplished by a single (judicious) substitution, namely the composition of the separate substitutions.

For both integrals in $(*)$, the choice $u = -t^2$ is the natural one, but only in the second case is the substitution immediately helpful. As it turns out, the first integral does not have an elementary antiderivative; informally, "e^{-t^2} cannot be integrated in closed form."

Trigonometric Integrals

Each of the trig functions has an elementary antiderivative. Clearly

$$\int^x \sin\theta\,d\theta = -\cos x + C, \qquad \int^x \cos\theta\,d\theta = \sin x + C,$$

while we found the antiderivative of tan above; cot is entirely similar. The secant function is, by comparison, difficult. For now, we merely state the formula, which may be checked by hand:

$$\int^x \sec\theta\, d\theta = \log\left|\sec x + \tan x\right| + C.$$

The derivative formulas for tan and sec give useful integrals:

$$\int^x \sec^2\theta\, d\theta = \tan x + C, \qquad \int^x \sec\theta\tan\theta\, d\theta = \sec x + C.$$

Finally, the double angle formulas

$$\cos^2\theta = \tfrac{1}{2}(1+\cos 2\theta), \qquad \sin^2\theta = \tfrac{1}{2}(1-\cos 2\theta)$$

can be antidifferentiated, yielding

$$\int^x \cos^2\theta\, d\theta = \tfrac{1}{4}(2x+\sin 2\theta)+C, \qquad \int^x \sin^2\theta\, d\theta = \tfrac{1}{4}(2x-\sin 2\theta)+C.$$

Integration by Parts

Integration by parts is the analogue of the product rule for differentiation. In Leibniz notation, if u and v are differentiable functions, then

$$d(uv) = u\,dv + v\,du, \qquad \text{or} \quad u\,dv = d(uv) - v\,du.$$

If the derivatives are continuous, that is, if u and v are $\mathscr{C}^1$ functions on $[a,b]$, then the previous equation may be integrated, yielding the *integration by parts* formula:

$$\int_a^b uv' = (uv)\Big|_a^b - \int_a^b vu'. \tag{15.5}$$

In order to apply integration by parts, the integrand in question must be written as a product of two functions, u and v', such that u' is easy to calculate, v is easy to find, and vu' can be antidifferentiated more easily that uv'. These are stringent requirements, but there are classes of functions for which integration by parts works well. There is an additional art in choosing u and v'. As above, example is the best illustrator.

- $u = t$ and $dv = \sin t\,dt$, so $du = dt$ and $v = -\cos t$.

$$\int^x t \sin t\,dt = -t\cos t\Big|^x + \int^x \cos t\,dt = \sin x - x\cos x + C.$$

Analogous choices of u and v' handle the integrands $t\cos t\,dt$ and $te^t\,dt$. If higher powers of t are involved, integration by parts yields a recursion formula. Taking u to be a power of t and dv to be a trig function,

$$\begin{aligned}\int^x t^n \sin t\,dt &= -t^n\cos t\Big|^x + n\int^x t^{n-1}\cos t\,dt \\ &= \big(-t^n\cos t + nt^{n-1}\sin t\big)\Big|^x - n(n-1)\int^x t^{n-2}\sin t\,dt \\ &= \big(-x^n\cos x + nx^{n-1}\sin x\big) - n(n-1)\int^x t^{n-2}\sin t\,dt\end{aligned}$$

This formula can be applied recursively without further work.

- $u = \log t$, $dv = t^n\,dt$.

$$\begin{aligned}\int^x t^n\log t\,dt &= \frac{1}{n+1}t^{n+1}\log t\Big|^x - \frac{1}{n+1}\int^x t^n\,dt \\ &= \frac{x^{n+1}}{(n+1)^2}\big((n+1)\log x - 1\big)\end{aligned}$$

Integration by parts can be used to express an integral in terms of itself in a non-trivial way. The next two examples and Exercise 15.4 are typical.

- $u = \sin^{n-1} t$, $dv = \sin t\,dt$,

$$\begin{aligned}\int^x \sin^n t\,dt &= -\cos x\,\sin^{n-1}x + (n-1)\int^x \sin^{n-2}t\,\cos^2 t\,dt \\ &= -\cos x\,\sin^{n-1}x + (n-1)\int^x \sin^{n-2}t\,dt \\ &\qquad - (n-1)\int^x \sin^n t\,dt.\end{aligned}$$

The unknown integral appears on both sides, and we may isolate it algebraically:

$$\int^x \sin^n t\,dt = -\frac{1}{n}\cos x\,\sin^{n-1}x + \frac{n-1}{n}\int^x \sin^{n-2}t\,dt. \tag{15.6}$$

This formula reduces the integral of $\sin^n t$ to the integral of $\sin^{n-2} t$.

- $u = \tan t$, $v' = \sec t \tan t \, dt$.

$$\begin{aligned}\int^x \sec t \tan^2 t \, dt &= \sec t \tan t \Big|^x - \int^x \sec^3 t \, dt \\ &= \sec t \tan t \Big|^x - \int^x \sec t(1 + \tan^2 t) \, dt \\ &= \sec x \tan x - \log \big| \sec x + \tan x \big| - \int^x \sec t \tan^2 t \, dt.\end{aligned}$$

Solving for the unknown integral,

$$\int^x \sec t \tan^2 t \, dt = \frac{1}{2}\Big(\sec x \tan x - \log \big| \sec x + \tan x \big|\Big) + C.$$

Trigonometric Substitution

The identity $\cos^2 \theta + \sin^2 \theta = 1$ and its variant $1 + \tan^2 \theta = \sec^2 \theta$ can be used to antidifferentiate many functions containing square roots. Let $a > 0$.

Integrand Contains	Substitution	Identity Utilized
$\sqrt{a^2 - x^2}$	$x = a \sin \theta$	$1 - \sin^2 \theta = \cos^2 \theta$
$\sqrt{a^2 + x^2}$	$x = a \tan \theta$	$1 + \tan^2 \theta = \sec^2 \theta$
$\sqrt{x^2 - a^2}$	$x = a \sec \theta$	$\sec^2 \theta - 1 = \tan^2 \theta$

It is equally possible to use the identities $\cosh^2 t - \sinh^2 t = 1$ and $1 - \tanh^2 t = \operatorname{sech}^2 t$ for these integrands if there is some reason to express the result in exponential form rather than trigonometric form.

Perusal of Chapter 13 will reveal that the following formulas are essentially definitions:

$$\int^x \frac{dt}{\sqrt{a^2 - t^2}} = \operatorname{Sin}^{-1} \frac{x}{a} + C \qquad \int^x \frac{dt}{a^2 + t^2} = \frac{1}{a} \operatorname{Tan}^{-1} \frac{x}{a} + C$$

$$\int^x \frac{dt}{t\sqrt{t^2 - a^2}} = \frac{1}{a} \operatorname{Sec}^{-1} \frac{x}{a} + C$$

Other integrals are transformed into a form we have found already.

- $t = a\sin\theta$, $dt = a\cos\theta\,d\theta$.

$$\int^{x} \bigl(\sqrt{a^2 - t^2}\bigr)^n\,dt = a^{n-1}\int^{\mathrm{Sin}^{-1}(x/a)} \cos^{n+1}\theta\,d\theta.$$

  This is handled with a reduction formula.

- $t = a\tan\theta$, $dt = a\sec^2\theta\,d\theta$.

$$\begin{aligned}\int^{x} \sqrt{a^2 + t^2}\,dt &= a^2\int^{\mathrm{Sec}^{-1}(x/a)} \sec^3\theta\,d\theta \\ &= a^2\int^{\mathrm{Sec}^{-1}(x/a)} \sec\theta(1+\tan^2\theta)\,d\theta \\ &= \frac{a^2}{2}\Bigl(\sec\theta\tan\theta + \log\bigl|\sec\theta+\tan\theta\bigr|\Bigr)\Bigr|^{\mathrm{Sec}^{-1}(x/a)} \\ &= \frac{a}{2}\Bigl(x\sqrt{x^2-a^2} + a\log\bigl|x+\sqrt{x^2-a^2}\bigr|\Bigr) + C.\end{aligned}$$

  The details are left to you, Exercise 15.5.

When converting back to the original variable after a trig substitution, the "right triangle trick" of Exercise 13.9 is helpful.

Partial Fractions

Every rational function with real coefficients can be written as a sum of terms of a certain standard form. This standard form, called a *partial fractions decomposition*, will be introduced in stages. Concretely, we wish to systematize identities such as

$$\frac{1}{1-x^2} = \frac{1}{2}\left(\frac{1}{1-x} + \frac{1}{1+x}\right), \qquad \frac{x}{1-x^2} = \frac{1}{2}\left(\frac{1}{1-x} - \frac{1}{1+x}\right),$$

or

$$\frac{2+2x+4x^2+x^3+x^4}{(1+x^2)^2(1+x)} = \frac{x}{(1+x^2)^2} + \frac{1}{1+x^2} + \frac{1}{1+x}.$$

Each term of a partial fractions decomposition is either a polynomial, or else a rational function whose denominator is a power of either a linear or an irreducible quadratic polynomial and whose numerator is a polynomial of degree less than the degree of the factor in the denominator. Symbolically, a partial fractions summand is of the form

$$\frac{b}{(x-c)^k}, \quad \text{or} \qquad \frac{ax+b}{(x^2+cx+d)^k}, \quad a,\ b,\ c,\ d \text{ real},\ c^2-4d<0.$$

Corollary 15.6 is instrumental in proving existence of a partial fractions decomposition: Every polynomial with real coefficients can be factored into a product of powers of irreducible linear and quadratic polynomials with real coefficients. By absorbing constants in the numerator, we may assume the irreducible factors of the denominator are monic.

Imagine the problem of trying to write the specific rational function $f(x) = x/(1-x^2)$ in partial fractions form. We will walk through the process, then re-examine the construction in general terms. The first step is to factor the denominator:

$$f(x) = \frac{x}{(1-x)(1+x)}.$$

Next pick one of the irreducible factors of the denominator, say $1-x$, and look for a constant b such that

$$\left(\frac{x}{(1-x)(1+x)} - \frac{b}{1-x}\right) = \frac{-b+(1-b)x}{(1-x)(1+x)}$$

has a finite limit at $x = 1$. Since the denominator vanishes at 1, the numerator must also vanish at $x = 1$ if the quotient is to have a limit. Thus $1-2b = 0$, or $b = 1/2$. Upon substituting this value of b into the right-hand side and simplifying, we find that our fraction has become simpler: The "new" denominator is a product of irreducible factors found in the original denominator, but the exponent of the factor we chose, $1-x$ in this case, has been reduced by at least one. Repeat the process; eventually the denominator becomes 1, since each step reduces the total degree of the denominator. In the present example, we are finished after one step.

It should be plausible that such an argument works in general, and a formal proof by induction on the degree of the denominator is straightforward. Intuitively, peeling off a partial fractions summand amounts to subtracting off "the highest order infinity" corresponding to a factor of the denominator, leaving an infinity of lower order. In lieu of a precise theorem (which is lengthy to state), here is a representative example: The partial fractions decomposition theorem guarantees that for each polynomial p of degree at most 6 (smaller than the degree of the denominator below), there exist constants a_i and b_i such that

$$\frac{p(x)}{(x^2+x+2)^2(x-3)^3} = \frac{a_1x+b_1}{(x^2+x+2)^2} + \frac{a_2x+b_2}{(x^2+x+2)} + \frac{b_3}{(x-3)^3} + \frac{b_4}{(x-3)^2} + \frac{b_5}{(x-3)}.$$

The relevance to elementary integration is this: If we can show that every partial fractions summand has an elementary antiderivative, then we will have shown that every rational function has an elementary antiderivative.

A term of the form $b/(x-c)^k$ has antiderivative

$$\frac{b}{(1-k)(x-c)^{k-1}} \quad \text{or} \quad b\log(x-c),$$

according to whether $k \neq 1$ or $k = 1$. If we are willing to work with rational functions having complex coefficients, we are done, since over $\mathbf{C}$ every rational function is reduced to a sum of such terms. However, it is desirable to find real antiderivatives of real functions, so we are forced to consider terms of the form $(ax+b)/(x^2+cx+d)^k$ with $c^2 - 4d < 0$.

The first simplification is to complete the square in the denominator. Setting $u = x + c/2$, we see that $x^2 + cx + d = u^2 + \alpha^2$ for some real α. We are therefore reduced to antidifferentiating

$$\frac{au+b}{(u^2+\alpha^2)^k}, \qquad \text{or} \quad \frac{u}{(u^2+\alpha^2)^k}, \text{ and } \frac{1}{(u^2+\alpha^2)^k}.$$

Note that the constants a and b are being used "generically"; they do not necessarily stand for the same numbers from line to line.

A term of the form $u/(u^2+\alpha^2)^k$ is handled by the substitution $v = u^2 + \alpha^2$, $dv = 2u\,du$. A term of the form $1/(u^2+\alpha^2)^k$ is handled with the trig substitution $u = \alpha\tan\theta$, which leads to

$$\int^x \frac{du}{(u^2+\alpha^2)^k} = \int^{\mathrm{Tan}^{-1}(u/\alpha)} \frac{\alpha\sec^2\theta\,d\theta}{\alpha^{2k}\sec^{2k}\theta} = \alpha^{1-2k}\int^{\mathrm{Tan}^{-1}(u/\alpha)} \cos^{2k-2}\theta\,d\theta.$$

This integral is handled by repeated integration by parts. At last we have tracked down all contingencies. To summarize:

Theorem 15.8. *If $f : (a,b) \to \mathbf{R}$ is a rational function, then there exists an elementary function $F : (a,b) \to \mathbf{R}$ such that $F' = f$.*

Note that F itself may not be rational. A further trigonometric substitution is used to deduce the following, which roughly asserts that "every rational function of sin and cos has an elementary antiderivative."

Corollary 15.9. *If R is a rational function of two variables, then*

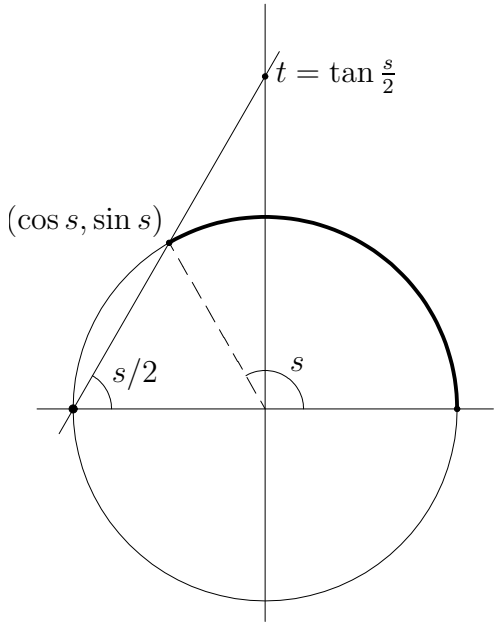
$$\int^x R(\cos s, \sin s)\,ds$$

is elementary.

Proof. (Sketch) Under the substitution $t = \tan\frac{s}{2}$, we have

$$ds = \frac{2\,dt}{t^2+1}, \qquad \cos s = \frac{t^2-1}{t^2+1}, \qquad \sin s = \frac{2t}{t^2+1},$$

so the integrand $R(\cos s, \sin s)\,ds$ becomes a rational function of t. One is led to this remarkable substitution by stereographic projection:



The details are left to you, Exercise 15.6. □

Existence and Non-existence Theorems

As repeatedly emphasized already, antidifferentiation is an art rather than a well-defined procedure. It is relatively difficult, even for someone with considerable experience, to tell at a glance whether or not a specific elementary function can be symbolically antidifferentiated. As you may have noticed, this section contains *ad hoc* techniques, dirty tricks, and hindsight reasoning. For example, several trigonometric integrals were evaluated by examining a table of derivatives and working backward. Other integrals are handled by trial-and-error, with various substitutions and integrations by parts. This state of affairs represents the nature of the subject. Computers are well-suited to this kind of "search" calculation, which at least partly explains their increasing popularity as tools for integration.

To complete the discussion, we state without proof two theorems regarding (non-)existence of elementary antiderivatives. The proofs rely on the concept of a "differential field", a field $\mathbf{F}$ equipped with a mapping $d : \mathbf{F} \to \mathbf{F}$ that satisfies the formal product rule: $d(ab) = a\,db + b\,da$. An example is the field of rational functions in one variable, with d the ordinary derivative. The general idea is to study when $df = g$ has a solution for given g, though the details are a little involved, and require more "abstract algebra" than we have developed. We have chosen the versions below because the statements are easy to understand, and because they lead to well-known examples.

Theorem 15.10. *Let f_1 and f_2 be rational functions of one variable. If the function*

$$g(x) = f_1(x)e^{f_2(x)},$$

has an elementary antiderivative G, then $G(x) = F(x)e^{f_2(x)}$ for some rational *function F.*

Theorem 15.10 implies that the following are not elementary:

$$\int^x e^{-t^2}\,dt, \qquad \int^x \frac{\sin t}{t}\,dt, \qquad \int^x \frac{dt}{\log t}.$$

The first integral is closely related to the "error function", which arises in probability. We investigate this integral in more detail below. To see that the latter two are not elementary, we consider the function $g(t) = e^t/t$, for which $f_2(t) = t$. By Theorem 15.10,

$$G(x) = \int^x \frac{e^t}{t}\,dt \quad \text{is elementary iff} \quad F(x) = \frac{1}{e^x}\int^x \frac{e^t}{t}\,dt \quad \text{is rational.}$$

However, a series calculation shows that $F(x) = \log x + O(1)$ near $x = 0$, so F is not rational. It follows that

$$\int^x \frac{\sin t}{t}\,dt = \frac{1}{2i}\int^x \frac{e^{it} - e^{-it}}{t}\,dt$$

is not elementary. The substitution $t = \log u$ shows that $\int^x du/\log u$ is not elementary, either.

Our final result about elementary antiderivatives is due to Chebyshev:

Theorem 15.11. *Let a, b, p, q, and r be real numbers. The antiderivative*

$$\int^x t^p(a+bt^r)^q\,dt$$

is elementary iff at least one of $\frac{p+1}{r}$, q, or $\frac{p+1}{r}+q$ is an integer.

For instance,

$$\int^x \sqrt{1+t^4}\,dt \quad \text{and} \quad \int^x t^2\sqrt{1+t^4}\,dt$$

are not elementary, while $\displaystyle\int^x t\sqrt{1+t^4}\,dt$ and $\displaystyle\int^x t^3\sqrt{1+t^4}\,dt$ are.

Definite Integrals

This section presents a miscellany of definite integrals and applications to evaluation of sums. In many cases, calculations are only sketched, and the details are left as exercises.

Proposition 15.12. *If n is a positive integer, then*

$$\int_0^{\pi/2} \sin^n t\,dt = \begin{cases} \frac{(n-1)}{n}\frac{(n-3)}{(n-2)}\cdots\frac{2}{3} & \text{if } n \text{ is odd} \\ \\ \frac{(n-1)}{n}\frac{(n-3)}{(n-2)}\cdots\frac{1}{2}\,\frac{\pi}{4} & \text{if } n \text{ is even} \end{cases}$$

In other words,

$$\int_0^{\pi/2} \sin^{2k+1} t\,dt = \frac{(2^k k!)^2}{(2k+1)!}, \qquad \int_0^{\pi/2} \sin^{2k} t\,dt = \frac{(2k)!}{(2^k k!)^2}\,\frac{\pi}{4}.$$

The substitution $u = (\pi/2) - t$ converts this integral into the integral (over the same interval) of a power of cosine.

Proof. The base cases are

$$\int_0^{\pi/2} \sin t\,dt = -\cos t\Big|_{t=0}^{\pi/2} = 1$$
$$\int_0^{\pi/2} \sin^2 t\,dt = \frac{1}{2}\int_0^{\pi/2}(1-\cos 2t)\,dt = \frac{1}{2}\Big(t-\frac{1}{2}\sin 2t\Big)\Big|_{t=0}^{\pi/2} = \frac{\pi}{4}.$$

The proposition follows from (15.6) and induction on n. □

The Gamma Function

Exercise 12.14 introduced the integral

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt, \qquad 0 < x < \infty, \tag{15.7}$$

which satisfies $\Gamma(n+1) = n!$ by induction on n. The integral is defined for non-integer x, and is an extension of the factorial function to the positive reals. This section develops a few properties of the Γ function, following Rudin.

The first issue is to establish convergence of the improper integral for the stated x. We split the integral into $\int_0^1 + \int_1^\infty$ and consider the pieces separately.

For every real α, $\lim_{t\to+\infty} t^\alpha e^{-t/2} = 0$ by Exercise 9.21. Thus, for all $x > 0$,

$$0 \le t^{x-1}e^{-t} \le \underbrace{(t^{x-1}e^{-t/2})}_{\to 0 \text{ as } t\to\infty} \cdot e^{-t/2} \le e^{-t/2} \qquad \text{for } t \gg 1.$$

Consequently, $\int_1^\infty t^{x-1}e^{-t}\,dt$ converges for all $x > 0$. Now, if $0 < x < 1$, then $0 \le t^{x-1}e^{-t} \le t^{x-1}$, so

$$\int_0^1 t^{x-1}e^{-t}\,dt < \int_0^1 \frac{1}{t^{1-x}}\,dt,$$

converges. If $x \le 0$, these integrals diverge, so the expression (15.7) is undefined.

The main result of this section is a characterization, due to Bohr and Mollerup, of Γ in simple abstract terms.

Theorem 15.13. *The function Γ satisfies the following properties:*

(a) $\Gamma(x+1) = x\Gamma(x)$ *for all* $x > 0$.

(b) $\Gamma(n+1) = n!$ *for all integers* $n > 0$.

(c) $\log\Gamma$ *is convex.*

Conversely, if $f : (0, +\infty) \to \mathbf{R}$ satisfies these properties, then $f = \Gamma$.

Proof. To establish (a), integrate by parts, using $u = t^x$ and $dv = e^{-t}\,dt$:

$$\begin{aligned}\Gamma(x+1) &= \int_0^\infty t^x e^{-t}\,dt = -t^x e^{-t}\Big|_{t=0}^{t=\infty} + \int_0^\infty x t^{x-1} e^{-t}\,dt \\ &= x\int_0^\infty t^{x-1}e^{-t}\,dt = x\Gamma(x).\end{aligned}$$

For (b), we compute that

$$\Gamma(1) = \lim_{b\to\infty}\int_0^b e^{-t}\,dt = \lim_{b\to\infty} -e^{-t}\Big|_{t=0}^{t=b} = \lim_{b\to\infty}(1-e^{-b}) = 1.$$

Induction on n guarantees $\Gamma(n+1) = n!$ for $n > 0$. To prove (c), observe that if $\frac{1}{p}+\frac{1}{q} = 1$, then the integrand of $\Gamma\big(\frac{x}{p}+\frac{y}{q}\big)$ is

$$t^{\frac{x}{p}+\frac{y}{q}-1}e^{-t} = t^{\frac{x}{p}+\frac{y}{q}-\frac{1}{p}-\frac{1}{q}}e^{-t(\frac{1}{p}+\frac{1}{q})} = \big(t^{x-1}e^{-t}\big)^{1/p}\big(t^{y-1}e^{-t}\big)^{1/q}.$$

Since $t^{x-1}e^{-t} > 0$ for $t > 0$, Hölder's inequality (Exercise 9.19) implies

$$\Gamma\big(\tfrac{x}{p}+\tfrac{y}{q}\big) \le \Big(\int_0^\infty t^{x-1}e^{-t}\,dt\Big)^{1/p}\Big(\int_0^\infty t^{y-1}e^{-t}\,dt\Big)^{1/q} = \Gamma(x)^{1/p}\Gamma(y)^{1/q}.$$

Taking logarithms,

$$\log\Gamma\big(\tfrac{x}{p}+\tfrac{y}{q}\big) \le \tfrac{1}{p}\log\Gamma(x) + \tfrac{1}{q}\log\Gamma(y),$$

which is the desired convexity statement.

Conversely, suppose $f : (0,+\infty) \to \mathbf{R}$ satisfies (a), (b), and (c), and set $\varphi = \log f$. Property (a) says $\varphi(x+1) = \varphi(x) + \log x$ for $x > 0$, and induction on n gives

$$\varphi(x+n+1) = \varphi(x) + \log\big((n+x)(n-1+x)\cdots(1+x)x\big). \tag{15.8}$$

Property (b) says $\varphi(n+1) = \log(n!)$ for every positive integer n and (c) says φ is convex.

Fix n, let $0 < x < 1$, and consider the difference quotients of φ on the intervals $[n, n+1]$, $[n+1, n+1+x]$, and $[n+1, n+2]$. Convexity of φ implies these difference quotients are listed in non-decreasing order. Property (a) implies that the difference quotient of φ on $[y, y+1]$ is $\log y$, so

$$\log n \le \frac{\varphi(n+1+x) - \varphi(n+1)}{x} \le \log n + 1. \tag{15.9}$$

Multiplying by x and substituting (15.8),

$$x\log n \le \varphi(x) + \log\big((n+x)(n-1+x)\cdots x\big) - \log(n!) \le x\log(n+1),$$

or

$$0 \le \varphi(x) - \log\frac{n!\,n^x}{x(x+1)\cdots(x+n-1)(x+n)} \le x\log\Big(1+\frac{1}{n}\Big) \tag{15.10}$$

As $n \to +\infty$, the upper bound goes to 0, so the squeeze theorem gives

$$\varphi(x) = \lim_{n\to\infty} \log\frac{n!\,n^x}{x(x+1)\cdots(x+n-1)(x+n)} \tag{15.11}$$

for $0 < x < 1$. Property (a) implies φ satisfies (15.11) for all $x > 0$.

Equation (15.11) says there is a unique function f satisfying properties (a)–(c), and that $\log f$ is the limit on the right; since Γ satisfies (a)–(c), the limit on the right must be equal to $\log\Gamma(x)$. □

An unexpected benefit of the argument is the equation

$$\Gamma(x) = \lim_{n\to\infty} \frac{n!\,n^x}{x(x+1)\cdots(x+n-1)(x+n)} \qquad \text{for } x > 0,$$

analogous to the characterization of exp as the limit of geometric growth:

$$e^x = \lim_{n\to\infty}\Big(1+\frac{x}{n}\Big)^n.$$

As in the case of exp, judicious application of this characterization of Γ leads to interesting identities. For example, define

$$\beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt \qquad \text{for } x,\, y > 0. \tag{15.12}$$

Theorem 15.14. *$\beta(x,y) = \dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ for all $x,\, y > 0$.*

Proof. Direct calculation gives $\beta(1,y) = 1/y$ for $y > 0$. Further, $\log\beta(\,\cdot\,,y)$ is convex in the sense that if $\frac{1}{p}+\frac{1}{q}=1$, then

$$\beta\Big(\frac{x}{p}+\frac{z}{q},y\Big) \le \beta(x,y)^{1/p}\beta(z,y)^{1/q} \qquad \text{for all } x,\, y,\, z > 0.$$

The proof is entirely similar to log convexity of Γ. Finally,

$$\beta(x+1,y) = \int_0^1 t^x(1-t)^{y-1}\,dt = \int_0^1 \Big(\frac{t}{1-t}\Big)^x(1-t)^{x+y-1}\,dt.$$

Integrating by parts, using $u = \big(t/(1-t)\big)^x$ and $v' = (1-t)^{x+y-1}$, gives

$$\beta(x+1,y) = -\frac{1}{x+y}\Big(\frac{t}{1-t}\Big)^x(1-t)^{x+y}\Big|_{t=0}^{t=1} + \frac{x}{x+y}\int_0^1 t^{x-1}(1-t)^{y-1}\,dt.$$

Since the "boundary term" is zero, $\beta(x+1,y) = \frac{x}{x+y}\beta(x,y)$. Putting these pieces together, the function

$$f(x) := \frac{\Gamma(x+y)}{\Gamma(y)}\beta(x,y)$$

satisfies properties (a)–(c) of Theorem 15.13, so $f = \Gamma$. □

Corollary 15.15. $2\displaystyle\int_0^{\pi/2} (\sin\theta)^{2x-1}(\cos\theta)^{2y-1}\,d\theta = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$.

This follows from the substitution $t = \sin^2\theta$. An interesting definite integral results from $x = y = 1/2$:

$$\pi = 2\int_0^{\pi/2} d\theta = \frac{\Gamma(\frac{1}{2})^2}{\Gamma(1)},$$

or

$$\sqrt{\pi} = \Gamma(\tfrac{1}{2}) = \textstyle\int_0^\infty t^{-1/2}e^{-t}\,dt.$$

The substitution $u = \sqrt{t}$, $du = dt/2\sqrt{t}$ yields

$$\int_0^\infty e^{-u^2}\,du = \frac{\sqrt{\pi}}{2}, \qquad \text{or} \qquad \int_{-\infty}^{+\infty} e^{-u^2}\,du = \sqrt{\pi}. \tag{15.13}$$

Taking $x = 1/2$ or $y = 1/2$ expresses the integrals of powers of sin and cos in terms of Γ, cf. Proposition 15.12.

The Error Function

Let $\mu \in \mathbf{R}$, $\sigma > 0$. The function

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma}}e^{-(x-\mu)^2/2\sigma}$$

arises naturally in probability as the famous *Gaussian density* or "bell curve". The variable x represents the result of a measurement of some sort, and the probability that the measurement lies between a and b is

$$P(a \le x \le b) = \frac{1}{\sqrt{2\pi\sigma}}\int_a^b e^{-(x-\mu)^2/2\sigma}\,dx$$

The *expected value* and *variance* are defined to be

$$E = \frac{1}{\sqrt{2\pi\sigma}} \int_a^b x e^{-(x-\mu)^2/2\sigma}\, dx, \quad V = \frac{1}{\sqrt{2\pi\sigma}} \int_a^b (x-\mu)^2 e^{-(x-\mu)^2/2\sigma}\, dx.$$

In terms of probability, repeated measurements of x are expected to cluster around E, and the average distance from a measurement to E is the *standard deviation* $\sqrt{V}$. The Gaussian above is normalized so that $E = \mu$ and $V = \sigma^2$, see Exercise 15.21.

Up to a linear change of variable, we may as well assume $\mu = 0$ and $\sigma = 1$. The *error function* $\operatorname{erf} : \mathbf{R} \to \mathbf{R}$ is defined by the improper integral

$$\operatorname{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt, \tag{15.14}$$

see Figure 15.5. Equation (15.13) implies that $\operatorname{erf}(x) \to 1$ as $x \to +\infty$. The probability that x is between a and b is $\operatorname{erf}(b) - \operatorname{erf}(a)$. Statistics

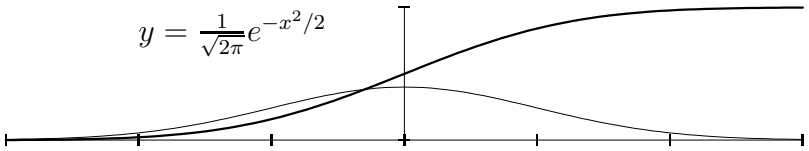


Figure 15.5: The Gaussian error function.

and probability textbooks often contain tables of erf. As noted above, erf is not an elementary function.

The Riemann Zeta Function

Consider the sum

$$\zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z}. \tag{15.15}$$

If we write $z = x + iy$ with x and y real, then the general summand has absolute value

$$|n^{-z}| = |e^{-z \log n}| = |e^{-x \log n} e^{-iy \log n}| = |e^{-x \log n}| = n^{-x}.$$

By comparison with the "p-series", the series (15.15) converges iff $x = \operatorname{Re} z > 1$, the convergence is absolute for all such z, and for every $\varepsilon > 0$,

the convergence is uniform (in z) on the set $\operatorname{Re} z > 1 + \varepsilon$. Though the series above does not converge when $\operatorname{Re} z \leq 1$, there is an extension to the "punctured" plane $\mathbf{C} \setminus \{1\}$ that is analytic in the sense that near each point $z_0 \neq 1$, there is a complex power series with positive radius whose sum is ζ. This analytic extension is the *Riemann zeta function*.

The Riemann zeta function, which is not elementary, is deeply pervasive throughout mathematics and physics. It is known (and not difficult to show) that the zeta function vanishes at the negative even integers, and that all other zeros lie in the "critical strip" $0 < \operatorname{Re} z < 1$. One of the outstanding open problems of mathematics is the *Riemann hypothesis*:

> Every zero of the Riemann zeta function that lies in the critical strip lies on the line $\operatorname{Re} z = 1/2$.

Anyone who makes substantial progress on this question will earn lasting fame; a complete resolution will garner mathematical immortality. As of May, 2000, a resolution of the Riemann hypothesis carries a prize of $1 million from the Clay Mathematics Institute.

Our extremely modest aim is to evaluate $\zeta(2)$. We separate the calculation into steps according to technique.

Lemma 15.16. $\displaystyle\int_0^1 \frac{\arcsin t}{\sqrt{1-t^2}}\,dt = \frac{\pi^2}{8}$

Proof. Use the substitution $u = \arcsin t$, taking appropriate care with the limit at $t = 1$. □

Lemma 15.17. $\displaystyle\int_0^1 \frac{t^{2k+1}}{\sqrt{1-t^2}}\,dt = \frac{(2^k k!)^2}{(2k+1)!}$

Proof. Use the substitution $t = \sin\theta$ and Proposition 15.12. □

Lemma 15.18. $\displaystyle\arcsin x = \sum_{k=0}^{\infty} \frac{(2k)!}{(2^k k!)^2}\frac{x^{2k+1}}{2k+1}$

Proof. According to Newton's binomial theorem (Theorem 14.18),

$$\frac{1}{\sqrt{1-x^2}} = 1 + \sum_{k=1}^{\infty} \frac{(2k)!}{(2^k k!)^2} x^{2k} \qquad \text{for } |x| < 1.$$

Integrate term by term, being careful with issues of non-uniform convergence near $x = 1$. □

Lemma 15.19. $\displaystyle\int_0^1 \frac{\arcsin t}{\sqrt{1-t^2}}\,dt = \sum_{k=0}^{\infty} \frac{1}{(2k+1)^2}$

Proof. Use Lemma 15.16 to expand the integrand:

$$\frac{\arcsin t}{\sqrt{1-t^2}} = \sum_{k=0}^{\infty} \frac{(2k)!}{(2k+1)(2^k k!)^2} \frac{t^{2k+1}}{\sqrt{1-t^2}},$$

then integrate term-by-term. Lemma 15.17 allows you to evaluate the resulting integrals, and a lot of cancellation occurs. □

Lemma 15.20. $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2} = \zeta(2) = \frac{\pi^2}{6}$

Proof. Separate the series into even and odd terms. (Why is this permissible?) You know the sum of the odd terms, and the sum of the even terms can be expressed using $\zeta(2)$ itself. □

There are systematic ways of evaluating the series $\zeta(2k)$ for k a positive integer, though better technology is required, such as complex contour integration or Fourier series. It turns out that $\zeta(2k)$ is an explicit rational multiple of π^{2k}. Interestingly, the values $\zeta(2k+1)$ are not known. In 1978, A. Apèry proved that $\zeta(3)$ is irrational.

Exercises

Exercise 15.1 The surface of the earth is divided into time zones that span roughly 15° of longitude, making a total time change of 24 hours as one circumnavigates the globe.[2] For this question, assume that everyone on earth keeps solar time: "Noon" is the time that the sun passes due south of you, assuming you're in the northern hemisphere. Assume that your longitude is 0°, and that it is noon for you. (Because this is mathematics, we're just re-defining "longitude"; you needn't travel to Greenwich!)

(a) Find a formula for the time at a point on the surface of the earth as a function of longitude. Be sure to adjust your time an angle units consistently.

[2] In reality, time zones obey political boundaries almost as much as geographic ones.

(b) Is the time of day a continuous function of position? If not, where is the discontinuity, and what happens to your measure of time if you cross the discontinuity?

(c) How is your time formula like the principle branch of the logarithm? According to your formula, what time is it at the north pole?

(d) Suppose there were a "solar time function" with no discontinuity. What would happen if you circumnavigated the globe?

In reality, the discontinuity is a fixed line in the Pacific Ocean rather than the "midnight point", which moves as the earth rotates. ◇

Exercise 15.2 Evaluate the following:

$$\int^x \frac{dt}{1-t^2} \qquad \int^x \frac{t\,dt}{1-t^2} \qquad \int^x \frac{dt}{t-t^2} \qquad \int^x \frac{dt}{t-t^3} \qquad \int^x \frac{1+t^2}{t-t^3}\,dt$$

Hint: Some of the partial fractions have been done for you. ◇

Exercise 15.3 Using the example in the text as a model, find recursion formulas for

$$\int^x t^n \cos t\,dt \qquad\qquad \int^x t^n e^t\,dt$$

Use a recursion formula to evaluate $\displaystyle\int^x t^4 \sin t\,dt$ ◇

Exercise 15.4 Find a recursion formula for $\displaystyle\int^x \cos^n t\,dt$ ◇

Exercise 15.5 Fill in the details of the evaluation of $\displaystyle\int^x \sec^3\theta\,d\theta$. ◇

Exercise 15.6 Fill in the missing details if the proof of Corollary 15.9; specifically, verify that under the substitution $t = \tan\frac{s}{2}$, the circular trig functions become rational functions of t. See Exercise 3.7 for details about stereographic projection. ◇

Exercise 15.7 Use the following outline to antidifferentiate sec: Multiply the numerator and denominator by cos, and use $\cos^2 = 1-\sin^2$. A substitution turns this into a rational function, whose partial fractions decomposition you know. ◇

Exercise 15.8 Use a theorem from the text to prove that $\displaystyle\int^x \sin t^2\,dt$ and $\displaystyle\int^x \cos t^2\,dt$ are not elementary. ◇

Exercise 15.9 Evaluate $\displaystyle\int_0^\infty e^{-t}\sin xt\,dt$ for x real. ◇

Exercise 15.10 Evaluate $\displaystyle\int_0^1 \frac{x^2+1}{x^4+1}\,dx$.
Hints: Divide the numerator and denominator by x^2, then use the (improper) substitution $u = x - 1/x$. ◇

Exercise 15.11 In this exercise you will evaluate $\displaystyle\int_0^{\pi/2} \log\sin x\,dx$

(a) Prove that the improper integral converges.
Suggestion: Look for suitable bounds on sin near 0.

(b) Show that $\displaystyle\int_0^{\pi/2} \log\sin x\,dx = \int_0^{\pi/2} \log\cos x\,dx$

(c) Use a substitution and the double angle formula for sin to evaluate the original integral.

If you can, find a way to evaluate the integral of log cos by a similar trick, without using part (b). ◇

Exercise 15.12 Prove that $\Gamma(n+\frac{1}{2}) = \dfrac{(2n-1)!!}{2^n}\sqrt{\pi}$. ◇

Exercise 15.13 Use formulas from the text to write the integrals

$$\int_0^{\pi/2} \sin^n\theta\,d\theta, \qquad \int_0^{\pi/2} \cos^n\theta\,d\theta,$$

in terms of the Γ function. ◇

Exercise 15.14 Evaluate

$$\int_{-1}^{1} (1-t)^n(1+t)^m\,dt, \qquad m,\ n \in \mathbf{N},$$

in terms of factorials.
Suggestion: First evaluate $\displaystyle\int_0^1 t^n(1-t)^m\,dt$ using the Γ function. ◇

Exercise 15.15 Prove that

$$\Gamma(x) = \frac{2^{x-1}}{\sqrt{\pi}}\Gamma(\tfrac{1}{2})\Gamma(\tfrac{x+1}{2})$$

for all $x > 0$. ◇

Exercise 15.16 Fill in the details of Lemma 15.16. ◇

Exercise 15.17 Fill in the details of Lemma 15.17. ◇

Exercise 15.18 Fill in the details of Lemma 15.18. ◇

Exercise 15.19 Fill in the details of Lemma 15.19. ◇

Exercise 15.20 Fill in the details of Lemma 15.20. ◇

Exercise 15.21 Let $\mu \in \mathbf{R}$ and $\sigma > 0$.

(a) Prove that

$$\frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} x e^{-(x-\mu)^2/2\sigma}\, dx = \mu.$$

Suggestion: Let $z = x - \mu$.

(b) Prove that

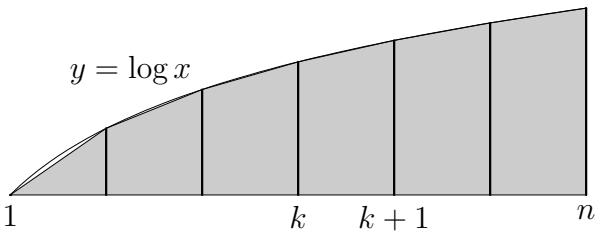
$$\frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma}\, dx = \sigma.$$

Suggestion: Use the substitution of part (a), then integrate by parts, using $u = z$.

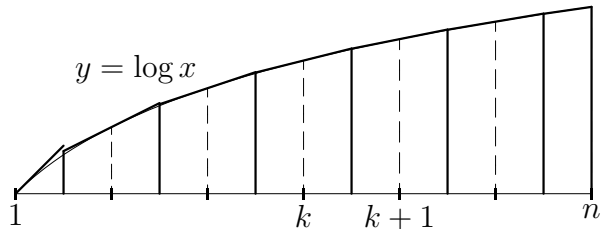
In each part, you may use the fact that $\lim(\mathrm{erf}, +\infty) = 1$. ◇

Exercise 15.22 This exercise presents a rough approximation to $n!$ for large n. The starting point is the observation that log is concave, so the graph lies below every tangent line and above every secant line. We will use these lines to get upper and lower bounds on the area under the graph of log between 1 and n.

(a) (Lower bound) Subdivide $[1, n]$ into $n - 1$ intervals of unit length. Find the total area of the trapezoids:



(b) (Upper bound) Consider the line tangent to the graph of log at the integer k. The area of the trapezoid lying below this line and between the vertical lines $x = k \pm \frac{1}{2}$ is an upper bound for the integral of log from k to $k+1$:



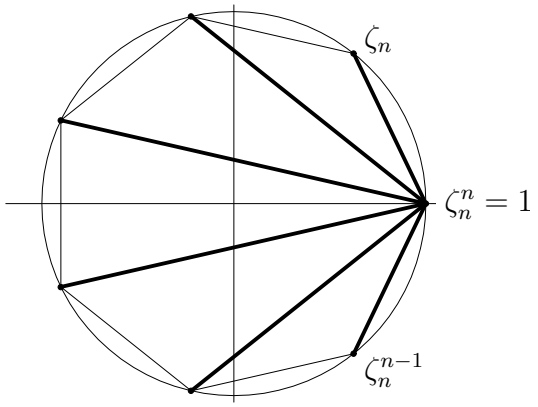
Show that the trapezoid centered at k, $1 < k < n$, has area $\log k$. When $k = 1$ or $k = n$, special considerations must be made. Find the total area of the trapezoids and triangle.

(c) Use the bounds found in parts (a) and (b) to prove that

$$e^{7/8}\frac{n^n\sqrt{n}}{e^n} < n! < e\frac{n^n\sqrt{n}}{e^n}.$$

In other words, $e^{7/8} < \dfrac{n!}{(n/e)^n\sqrt{n}} < e$ for $n \geq 1$. ◇

Exercise 15.23 Let $n \geq 3$ be an integer, and set $\zeta_n = e^{2\pi i/n}$; recall that the set of powers of ζ_n constitutes the vertices of a regular n-gon inscribed in the unit circle. Consider the set of chords joining $1 = \zeta_n^n$ to the other $n-1$ vertices.



Prove that the product of the lengths of these chords is n, the number of sides! (Only a few special cases can be computed by hand.) $\diamond$

Exercise 15.24 (The momentum transform) Let $\varphi : (-1, 1) \to \mathbf{R}$ be continuous and positive.

(a) Prove that the equation

$$x = \int_0^{u(x)} \frac{dt}{\varphi(t)}$$

defines an increasing, $\mathscr{C}^1$ function u, and that $u' = \varphi \circ u$. What are the domain and image of u?

(b) Prove that the equation

$$f(x) = \int_0^{u(x)} \frac{t\,dt}{\varphi(t)}$$

defines a $\mathscr{C}^2$ function f, and show that $f' = u$.

(c) Suppose $\varphi(t) = 1 - t^2$. Find u and f as explicit functions of x.

(d) Suppose $\varphi(t) = 1 + t^2$. Find u and f as explicit functions of x.

Hint: You will need techniques from throughout the book, but there is a reason this question was saved for the final chapter. $\diamond$

Postscript

In the landmark essay *The Cathedral and the Bazaar*, Open Source Software advocate Eric S. Raymond wrote, "Every good work of software starts by scratching a developer's personal itch." Whatever this book's quality as an instructional tool, it scratches an itch, namely the author's desire to present interesting, living, rigorous mathematics to students whose background is the 1990s high school mathematics curriculum. Many things have changed since the author was a high school student; today there is a greater emphasis on conceptual understanding without technical details, discovery through experimentation, numerical approximation, and applications.

While these are laudable goals, there is a danger in presenting mathematics as an experimental science, and in relying on intuitive principles and plausibility arguments instead of precise definitions and logical, deductive proof. In the realm of a calculus course the risks may not be so apparent, because problems have often been carefully chosen and stated to be amenable to the conceptual methods being taught. However, real life is never as clean and simple as a textbook, and pure intuition without technical knowledge easily goes astray. This book tries to bridge the worlds of technique and understanding, presenting mathematics as a language of precision that is guided by intuitive principles, and that leads to beautiful, unexpected destinations.

This book started as a set of course notes from Analysis I (MAT 157Y) taught in the year 1997–8 at the University of Toronto. While the course was not an unimprovable success by all measures, it was among the most satisfying extended teaching experiences I have had. A senior colleague at the "UofT" once said that mathematicians should advertise themselves to students as weight trainers of the mind.[1] In this metaphor, Analysis I was greatly successful. The problem sets were extremely challenging, but at least one student always made some

[1] "We are pumping brains!"

progress on a question. Students were spurred to investigate issues beyond what the homework asked, and were thereby led to deeper understanding and intellectual independence. A substantial fraction of the class went on to graduate programs in mathematics and physics at top research schools. My first debt of gratitude is to the students of Analysis I, particularly Jameel Al-Aidroos, Sunny Arkani-Hamed, Ari Brodsky, Chris Coward, Chris George, Fred Kuehn, Brian Lee, Cuong Nguyen, Caroline Pantofaru, Dave Serpa, and Dan Snaith.

MAT 157Y would also have been far less successful without the Herculean efforts of the teaching assistants, Gary Baumgartner and Blair Madore, who tirelessly fired the students' curiosity, creatively constructed examples and metaphors, and tempered my flights of abstraction both in the tutorials and when creating tests and exams. A couple of Gary's incisive supremum questions appear as exercises.

The nucleus of Chapter 1 was a handout by Steve Hook, on writing proofs, that he distributed to his real analysis course at UC Berkeley during the summer of 1986.

Andres del Junco at the University of Toronto kindly provided his course materials for Analysis I. Their influence is present at several places in the text, either as examples or as inspiration for exercises.

The proof of the Weierstrass approximation theorem comes from a one-hour undergraduate lecture given by Serge Lang at UC Berkeley in the late 1980s.

In a letter to the *Notices* of the American Mathematical Society, Donald Knuth suggested writing a calculus book based on O-notation. When I came across his letter—years after it was written, and after a non-negligible fraction of this book had been written—the prospect of "porting" the book from ε-δ language to O-notation seemed both doable and pedagogically well-advised. The "calculi of sloppiness" are both concrete enough to be psychologically satisfying to students and powerful enough to express truly subtle ideas. Whether or not this book can be considered a successful implementation of Knuth's vision remains to be seen.

I am grateful to several people who posted to the Usenet group `sci.math`, but especially to Matthew Wiener, whose exposition on integration in elementary terms was the impetus to include some general theorems in Chapter 15 rather than just scattered folklore results. John Baez's web page *Weekly finds in mathematical physics* was also a source of inspiration. His account of the Weierstrass $\wp$ function, in analogy with the definition of circular trig functions using integrals of algebraic

functions, was very tempting to include in Chapter 15, and was only omitted with regret.

Over the years I have collected interesting exercises, factoids, and other mathematical tidbits, often recorded on miscellaneous scraps of paper or in fallible memory, and I do not always know the origins of these items. I offer my sincerest apologies to anyone whose work is uncited.

This book was produced with Free software on a GNU/Linux system. The concept of "Free" software is important to academics, and is worth explaining briefly. A computer program is a set of instructions for performing some task, and (practically speaking) can be easily copied. Software is written as *source code* in a human-readable *programming language*, and is turned into machine-readable *executables* or *binary files* with a special program called a *compiler*. It is helpful to think of a recipe (instructions for creating a dish), which can be copied and shared without losing the original. In the early days of computing (before about 1980), software was written and shared like recipes. In the 1980s, an industry rose up around the concept of software as a commodity, like *ingredients* for a recipe. In this new model, sharing is forbidden, and the ingredients (source code) are secret. Unfortunately, because it is technically possible to copy software, it is necessary to treat all software users as potential thieves. The Free software movement aims to create a community in the spirit of software sharing, where everyone is at liberty to view and modify the source code of programs to suit their particular needs. The GNU project, started in 1984 by Richard M. Stallman, set itself the goal of creating an entire computer operating system from scratch and placing it under a license that would allow anyone to read, modify, and distribute the source code, subject only to the restriction that these terms of openness may not be revoked. Today, the GNU/Linux operating system is used worldwide by millions of people. "Linux" is Linus Torvalds' Unix-like *kernel*, the program that allocates hardware resources to all other running programs. Free software is arguably the only acceptable software in an academic setting; as a scientist or mathematician, you cannot fully trust the results of a computation unless you know exactly what the computer is doing. Free software is amenable to user inspection. While you may never read the source code for the `C` compiler or the Linux kernel, it is crucial that many knowledgeable people *have* audited the code, and that you (the individual user) retain the *right* to audit the code if you choose. In the most pragmatic sense, this right is no different from accurate

ingredient labelling on food. This book was produced with `emacs`, the GNU text editor, and teTEX, an implementation of Donald Knuth's Free TEX typesetting engine. The figures were created with `ePiX`, a Free utility for creation of mathematically accurate figures.

A few textbooks have had an obvious influence on this book. Most notable is *Calculus* (3rd edition) by Michael Spivak [8], the text for MAT 157Y. Despite the excellence of that text, and its suitability for a course like MAT 157Y, I felt the wish for a slightly different emphasis, arrangement, coverage, and style. Walter Rudin's classic *Principles of Mathematical Analysis* [6] was also very influential, both because I was first exposed to real analysis by this gem, and because Rudin has impeccable taste for choosing examples and exercises that highlight subtle technicalities that are invisible from a casual first inspection. The older calculus texts by Apostol [2], and Courant and John [3] were a source of inspiration, but (sadly) like Spivak's *Calculus* are seldom regarded as suitable for modern students. *Calculus: A Liberal Art* by William Priestley [5], is a delightful first course in calculus, but is aimed at Liberal Arts students who do not intend to pursue further studies in mathematics. Finally, the "Harvard" text, by Deborah Hughes-Hallet, Andrew Gleason, et. al., had a distinctive influence on the style of this book. This possibly surprising admission deserves lengthier comment.

In my experience, technical details are analogous to a skeleton, and conceptual intuition is analogous to muscles and skin. Without muscles, a skeleton is stiff and inert. Without a skeleton, muscles cannot move in a directed way. Together, acting in synergy, they grant us strength and graceful movement. In the middle of the 20th Century, the pendulum of mathematics education had swung toward extreme formalism (the New Math). Starting in the late 1980s, mathematics educators began to embrace a kind of non-technical conceptual understanding as a remedy to the "mindless formalism" that was failing to reach most students. This trend is in full force at present, and is widely apparent in the content and style of calculus texts of the early 21st Century. While "relevance" and "inclusiveness" are desirable goals for mathematical pedagogy, the effort is doomed to lose its intellectual substance if the rigorous foundations of mathematics are forgotten entirely. The debacle of algebraic geometry in the early 20th Century comes to mind: Theorems were proven by intuitive arguments, and were often incorrect. The literature of enumerative geometry was damaged, and development of the field delayed for decades until a proper foundation was built.

No one will be well-served if generations of students, many of them

future teachers, grow up without exposure to the technical details of real analysis. This book is a modest attempt to imbue the muscles of conceptual understanding of calculus with a skeleton of logical and technical formalism, that is, to embrace the educational trend of conceptual understanding without losing sight of the intellectual bedrock of mathematics. The foundations of calculus and its exposition were laid centuries ago, and it would be ludicrous to claim any credit of originality in material or presentation. Nonetheless, I believe this book fills a niche. While not every student is expected to read the book sequentially cover to cover, it is important to have the details in one place. Calculus is not a subject that can be learned in one pass. Indeed, this book nearly assumes readers have already had a year of calculus, as had the students of MAT 157Y. I hope this book will grow with its readers, remaining both readable and informative over multiple traversals, and that it provides a useful bridge between current calculus texts and more advanced real analysis texts.

Andrew D. Hwang
May 18, 2003
Sterling, MA

Bibliography

[1] Lars V. Ahlfors, *Complex Analysis*, 3rd edition, McGraw-Hill, 1979.

[2] Tom M. Apostol, *Calculus*, Blaisdell (Random House), 1962.

[3] Richard Courant and Fritz John, *Introduction to Calculus and Analysis*, Wiley Interscience, 1975.

[4] Serge Lang, *A First Course in Calculus*, 3rd edition, Addison-Wesley, 1973.

[5] William McGowen Priestley, *Calculus: A Liberal Art*, 2nd edition, Springer-Verlag, 1998.

[6] Walter Rudin, *Principles of Mathematical Analysis*, 3rd edition, McGraw-Hill, 1976.

[7] George F. Simmons, *Differential Equations with Applications and Historical Notes*, 2nd edition, McGraw-Hill, 1991.

[8] Michael D. Spivak, *Calculus*, 3rd edition, Publish or Perish, 1994.

Index